Smoothed Probabilistic-based Algorithms for Sparse Data with application to Emotion Recognition and Sentiment Analysis

Fatma Najar

A Ph.D. Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering (CIISE)

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Information and Systems Engineering) at

Concordia University

Montréal, Québec, Canada

August 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Fatma Najar**

Entitled: **Smoothed Probabilistic-based Algorithms for Sparse Data with application to Emotion Recognition and Sentiment Analysis**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Information and Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair
Dr. Rolf Wuthrich

______________________________ External Examiner
Dr. Antoine Tabbone

______________________________ External To Program
Dr. Lyes Kadem

______________________________ Examiner
Dr. Jamal Bentahar

______________________________ Examiner
Dr. Farnoosh Naderkhani

______________________________ Supervisor
Dr. Nizar Bouguila

Approved by ______________________________
Abdessamad Ben Hamza, Chair
Department of Concordia Institute for Information Systems Engineering (CIISE)

____________ 2022 ______________________________
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Smoothed Probabilistic-based Algorithms for Sparse Data with application to Emotion Recognition and Sentiment Analysis

Fatma Najar, Ph.D.

Concordia University, 2022

Humans are able to express more than 10,000 expressions through 43 facial muscles which makes reading faces a significant human skill and a challenge task for Artificial Intelligence (AI) algorithms. Even though much research work has been proposed for the field of sentiment analysis and emotion recognition, it continues to present considerable challenges. In our research, we focus on providing novel emotion recognition and sentiment analysis solutions where we address data challenges that occur in different modalities: texts, images, and videos. Considering these different multimedia contents, the analysis of data considers the concurrency nature of words in a collection of documents, visual words or proportional features vectors when considering images and videos. This type of data involves several challenges including sparseness, burstiness, correlated features, and high-dimensionality.

In this dissertation, we propose smoothed probabilistic-based approaches to deal with the aforementioned data challenges. First, we introduce the calculation of the exact Fisher information matrix of the generalized Dirichlet multinomial. Our proposed approach has been adopted for detecting depression in tweets, dialogue-based emotion recognition, and image-based sentiment analysis. Second, we develop different smoothed solutions for handling sparsity, high dimensionality, and burstiness issues such as smoothed Dirichlet multinomial, smoothed Generalized Dirichlet, smoothed Generalized Dirichlet multinomial (SGDM), Taylor approximation to the SGDM, Latent-based smoothed Beta-Liouville, Smoothed Beta-Liouville Emotion Term model, and Smoothed Scaled Dirichlet Relevance Model. These models are based on smoothing count vectors in a smoothed subset of the whole simplex to deal with the problem of sparseness. Moreover, we incorporate a hierarchical generalized Dirichlet prior for sparse multinomial distributions and a Beta-Liouville Naive Bayes with vocabulary knowledge. These two techniques build up on Bayesian vocabulary knowledge over large discrete domains represented by subsets of feasible outcomes: "observed" and "unobserved" words. In another research work, we consider a sparse topic model for non-exchangeable correlated data over time and present a new interactive distance dependant IBP compound Dirichlet process. We derive a Markov Chain Monte Carlo sampler combined with Metropolis-Hastings algorithm and study its performance on sentiment analysis data.

Acknowledgments

I would like to express my deepest thank for my supervisor Professor Nizar Bouguila. Thank you Dr. Nizar for all your continuous support, encouragements, advice, and for your unconditionally availability whenever needed. You vision, dynamism, and immense knowledge have always inspired me to keep going during this journey.

I would like to thank my committee members for their valuable time in reviewing my work, and for their valuable discussion and insightful comments throughout each milestone of my PhD.

I dedicate this thesis to my beloved mother and father, Sarra Baccar and Ridha Najar. Thank you for all the love you have passed to me. Mother, thank you for your endless care, for your unconditional love, for handling me and being my side during this journey, and for your prayers. Father, thank you for believing in me, for pushing me always to be the best version of myself, and for your continuous guidance during all my studies. My parents, nothing would be possible without your endless support and love. Thank you for everything, Love you. To my dear two sisters Khadija and Zeineb, you mean everything to me. Thanks for your caring, your love, and for being my lovely sisters.

I would like to express my deepest gratitude for the enjoyable moments that I spent in Concordia University, and in particular, our XAI Lab. Thank you my friends, lab mates, colleagues and research team for all the cherished time spent in the lab, and in social gatherings.

Special thanks to all my financial supports; this research wouldn't have been possible without Concordia University funds, Natural Sciences and Engineering Research Council of Canada (NSERC), and Mitacs.

“Optimism is the faith that leads to achievement.
Nothing can be done without hope and confidence.”
–Helen Keller–

Contributions of Authors

This Ph.D. Thesis consists of eight manuscripts. Five manuscripts have been published, and the rest have been submitted for publication in refereed academic journals. Each chapter consists of the content of a manuscript which has been reformatted and reorganized according to the requirements set out in the guideline by the School of Graduate Studies.

✻ **Manuscript 1 (Chapter 2)**: Fatma Najar and Nizar Bouguila, “Exact fisher information of generalized Dirichlet multinomial distribution for count data modeling”, Information Sciences, Vol. 586, pp. 688-703, March 2022.

✻ **Manuscript 2 (Chapter 3)**: Fatma Najar and Nizar Bouguila, “Emotion recognition: A smoothed Dirichlet multinomial solution”, Engineering Applications of Artificial Intelligence journal, Vol. 107, article 104542, January 2022.

✻ **Manuscript 3 (Chapter 4)**: Fatma Najar and Nizar Bouguila, “Smoothed Generalized Dirichlet: a novel count data model for detecting emotional states”. In IEEE Transactions on Artificial Intelligence, October 2021.

✻ **Manuscript 4 (Chapter 5)**: Fatma Najar and Nizar Bouguila., “Latent Smoothed Beta-Liouville Topic Modeling for Emotion Analysis and Affect Recognition. Submitted to IEEE Transactions on Emerging Topics in Computational Intelligence (2022).

✻ **Manuscript 5 (Chapter 6)**: Fatma Najar and Nizar Bouguila, “On smoothing and Scaling Language Model for Sentiment Based Information Retrieval”. Accepted with Minor revision in Advances in Data Analysis and Classification Journal.

- ✻ **Manuscript 6 (Chapter 7)**: Fatma Najar and Nizar Bouguila, "Sparse Generalized Dirichlet Prior Based Bayesian Multinomial Estimation". In International Conference on Advanced Data Mining and Applications (ADMA'21), pp. 177-191, Sydney, Australia, 2-4 February 2022.

- ✻ **Manuscript 7 (Chapter 7)**: Fatma Najar and Nizar Bouguila, "Sparse Document Analysis using Beta-Liouville Naive Bayes with Vocabulary Knowledge". In 16th International Conference on Document Analysis and Recognition (ICDAR 2021), Lausanne, Switzerland, September 5-10, 2021.

- ✻ **Manuscript 8 (Chapter 8)**: Fatma Najar and Nizar Bouguila, "Interactive Distance Dependent IBP Compound Dirichlet Process". Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement and Motivations

Emotions play a fundamental part in human experiences. Imagine how much could emotions shape the way we experience the world. Accordingly, humans express their emotions in multiple ways such as speech intonation, textual contents, facial expressions, body posture, and gestures. These factors create different types of variables in emotion recognition which alleviate the need of categorizing human emotions and analyzing sentiments from multimodal content. Sentiment analysis is the process of mining text data to identify and classify subjective information basically into positive, negative or neutral polarities while for emotion recognition, the psychological or mood states are detected to include objective information. A better understanding of emotions helps Artificial Intelligence (AI) makes life easier and more productive in various fields: customer service, product/market research, healthcare, automotive, education, and gaming. Therefore, significant efforts have focused on extracting and gathering information from social media which has been always considered as a double-edged sword and could be viewed as a first reason that affect teenager emotional states. The majority of people use social media to express their emotional states, happy moments, anxiety, and even sadness which give the opportunity to understand more how humans convey their psychological states. Consequently, it has become increasingly difficult for researchers to mine knowledge from these different modalities and to keep up with the considerable challenges of data modeling. The analysis of such information leads to “count data” which consider the concurrency nature of words in a collection of text documents, visual words or proportional features vectors when considering multimedia datasets. Hence, there are major challenges of count data which need to be overcome, e.g., sparsity, curse of dimensionality, correlated features, and burstiness. Sparseness, or the phenomenon of data with excess of zeros, appears whenever a word has not occurred in a document when considering a “bag-of-words” structure. The burstiness phenomenon [3] described as an accidental repetition of infrequent words or phrases in long documents is somehow a consequence of the overused conditional assumption of independence. Earlier research works have considered basic distributions such as the Poisson [4], Hurdle model [5], zero-inflated [6], Negative-Binomial [7], and multinomial response models [6, 8]. However, the Poisson distribution has only one parameter that induces the problem of overdispersion. Further, a central problem that rises with the multinomial distribution is the “Naive Bayes assumption” which imposes the independence of all the features and assumes that the occurrences of attributes are learned separately. Despite the accurate performance achieved with the multinomial distribution in several applications [9, 10] and the different research works proposed to deal with the multiple challenges of count data, they are very simple to be able to deal with emotion and sentiment analysis applications.

The goal of this thesis is to address the aforementioned problems through different smoothed probabilistic-based approaches. In this context, we detail in the next chapters the motivations that led us to present each proposed approach.

1.2 Thesis Contributions

In this thesis, we propose novel solutions for emotion recognition and sentiment analysis where we address count data challenges through the following contributions:

- **Exact Fisher Information of Generalized Dirichlet multinomial Distribution for count data modeling.**

  We tackle the problem of Naive Bayes assumption considered in Fisher information matrix. We propose an exact calculation of the Fisher information matrix (EFIM) for the Dirichlet multinomial [11] and generalized Dirichlet multinomial (GDM) mixture model where we introduce a new Deterministic-annealing EM learning algorithm based on Fisher scoring algorithm and minimum message length. We consider different modalities of count data (text, dialogue, image) generated from three challenging applications namely detecting depression in tweets, dialogue-based emotion recognition, and image-based sentiment analysis.

This work is published in *Information science* journal [12] and Canadian AI conference [11].

- **Emotion recognition: A smoothed Dirichlet multinomial solution.**

  We propose a new count-data model namely smoothed Dirichlet multinomial (SDM) mixtures with a likelihood-based learning. Based on the smoothed Dirichlet, we develop two other novel approaches: SD distribution-based word embedding and SD-based agglomerative hierarchical technique. These works are published in IRI [13] and ISVC [14]. Afterwards, we detect emotional states using SDM by means of challenging applications such as depression detection, psychology analysis, and pain estimation.

  This work is published in *Engineering Applications of Artificial Intelligence* journal [15].

- **Smoothed Generalized Dirichlet: a novel count data model for detecting emotional states.**

  We focus on tackling the sparseness and overdispersion problems by proposing a Smoothed Generalized Dirichlet distribution, the learning algorithm for estimating the parameters, and two clustering mechanisms namely mixture model based on EM algorithm and geometrical information using Kulback-Leibler, Fisher information, and Bhattacharyya distance. We present a new smoothed prior to the multinomial distribution to deal with burstiness and overdispersion problems; the Smoothed Generalized Dirichlet multinomial (SGDM); and a Newton-Raphson algorithm for learning the resulting model. For high-dimensional issue, we approximate the SGDM using Taylor series expansion (TSGDM). Next, we detect emotional states through two challenging applications: human pain intensity expressed through images/videos and disaster tweets related emotions.

  This work is published in *IEEE transaction on Artificial Intelligence* [16].

- **Latent Smoothed Beta-Liouville Topic Modeling for Emotion Analysis and Affect Recognition.**

We develop a new statistical approach based on Beta-Liouville distribution and a smoothed simplex: Latent-based Smoothed Beta-Liouville. We incorporate modeling unknown documents in a Bayesian folding-in way along with estimating the document-topic distribution to present the Smoothed Beta-Liouville kernels for PLSI. Using these two novel topic models, we track emotions in children-directed texts ”fairy tales” and we detect affective states from facial and body recognition through a bimodal affect framework that combines facial expressions features, pose estimation, and hand gestures. We consider the correlations between different behavioral motions with latent topics using Latent SBL for affect recognition from face and body.

This work is submitted to *IEEE Transactions on Emerging Topics in Computational Intelligence*.

- **On Smoothing and Scaling Language Model for Sentiment Based Information Retrieval.**

  Mainly concerned with sentiment analysis, we propose a smoothed probabilistic-based approach for information retrieval; Smoothed Scaled Dirichlet Relevance Model (SSD-RM) and introduce the maximum likelihood estimation of the parameters that will be used in the retrieval framework. We propose a feature generation from the information retrieval system instead of the bag-of-words structure to tackle the problem of sparsity and dimensionality. Then, we proposed a new sentiment analysis framework combining the SSD-RM and SVM by means of Kulback-Leibler divergence.

  This work is submitted to *Advances in Data Analysis and Classification* journal.

- **Sparse Generalized Dirichlet Prior based Bayesian multinomial estimation.**

  With the focus on large discrete domains, we propose a novel sparse generalized Dirichlet prior based Bayesian multinomial estimation. We define a new prior over exponential hypothesis using vocabulary knowledge; each of which represents a set of feasible outcomes for seen and unseen words. We predict emotions revealed in sparse dictionary of German and English poetry and we analyze the flow of emotions related to natural disasters.

  This work is published in *International Conference on Advanced Data Mining and Applications* [17].

- **Sparse Document Analysis using Beta-Liouville Naive Bayes with Vocabulary Knowledge.**

  Focusing on the nature of short texts and large-scale documents, we propose a novel Beta-Liouville hierarchical prior over the multinomial estimates of the Naive Bayes classifier. We incorporate vocabulary knowledge to analyze emotion intensity and detect hate speech tweets which are marked with sparseness of its data.

  This work is published in *International Conference on Document Analysis and Recognition* [18].

- **Interactive Distance Dependent IBP Compound Dirichlet Process.**

  We address sparsity in topic modeling, the dependency between features over time, and the exchangeability data assumption. Accordingly, we propose a Spike and Slab prior where we smooth topic-word and topic-document distributions by introducing Bernoulli variables over words and topics, respectively. We introduce, interactive distance dependent Indian Buffet compound Dirichlet process (idd-ICDP), a novel nonparametric Bayesian for the purpose of considering the non-exchangeability of the data through sampling topics/words assignments using the distance between them over time. We integrate human experts knowledge for the purpose of improving topic quality using an objective topic-word distribution. We consider the proposed model as a supervised topic model for sentiment analysis. This work is submitted to *IEEE Transactions on Knowledge and Data Engineering*.

1.3 Thesis Outline

The rest of the thesis is organized as follows:

- Chapter 2: presents the exact calculation of Fisher matrix for generalized Dirichlet multinomial.

- Chapter 3: introduces the new smoothed Dirichlet multinomial and its EM learning algorithm.

- Chapter 4: discusses the novel Smoothed generalized Dirichlet distribution with different learning methods and Taylor series approximation.

- Chapter 5: displays the Latent-based Smoothed Beta-Liouville and Smoothed Beta-Liouvile Emotion Term model.

- Chapter 6: introduces the Smoothed Scaled Dirichlet (SSD) prior, the smoothed Scaled Dirichlet Relevance Model and the sentiment analysis framework based on SSD.

- Chapter 7: presents the hierarchical generalized Dirichlet prior for sparse multinomial distributions and the Beta-Liouville Naive Bayes with vocabulary knowledge.

- Chapter 8: describes the interactive distance dependant IBP compound Dirichlet process.

- Chapter 9: concludes the thesis with discussion remarks and future insights.

Chapter 2

Exact Fisher Information of Generalized Dirichlet Multinomial Distribution for count data modeling

Despite the multiple benefits of the Fisher information matrix, it is generally disregarded and substituted by the identity matrix or an approximation format. However, when dealing with complicated real-world applications, ignoring the correlation between data features may compromise the modeling capability. To address this problem we present the exact calculation of the Fisher information matrix (EFIM) for the generalized Dirichlet multinomial (GDM) mixture that has proven its efficiency when modeling count data. We present a parametrization of GDM mixture model that allows the determination of the Fisher matrix's elements by means of the Beta-binomial probability function. We also propose a novel count data modeling approach with the benefit of EFIM. In particular, we tackle the problem of mixture model estimation and selection using the Fisher scoring algorithm and minimum message length within the Deterministic Annealing Expectation-Maximization learning framework. Experiments on detecting depression in tweets, dialogue-based emotion recognition, and image-based sentiment analysis confirm the capability of the proposed approach and the merits of using the EFIM as compared with existing state-of-the-art methods and techniques that ignore the full determination of the Fisher information matrix's elements.

2.1 Introduction

Count data, in statistical modeling, are discrete variables, non-integer values, that range from zero to infinity. “Count” as the nature of the verb implies, is the enumeration of certain events, units, or items according to observations from different disciplines [19]. Count data have been considered first in econometric research using basic count models such as Poisson and Negative Binomial [20]. Then, the use of count data was extremely broad, referring to almost everything that has been occurred in a fixed period of time, for instance, in finance, ecology, biomedical, machine learning and data mining applications.

Examples of applications where count data are naturally generated include text retrieval and image categorization. In these applications, texts are represented as vectors of frequencies of words, and images are represented as frequencies of visual keywords. Several researchers advocate the use of Poisson distribution for count data modeling. Yet, Poisson distribution has only one parameter that characterizes the mean and is equal to the covariance which gives rise to the equal dispersion criterion. Having data with an observed covariance greater than the one predicted causes the problem of overdispersion. In addition, Poisson distribution [4] is very simple to be able to deal with count data challenges (overdispersion, sparsity). There are actually other alternative models broadly accepted to describe count data including Hurdle models [5], zero-modified distributions (zero-inflated) [6, 8], and multinomial response models. Multinomial distribution has demonstrated its capability to analyze count data in several related applications [21, 10]. However, a central problem with this distribution is the so-called “Naïve Bayes assumption”. This hypothesis imposes the independence of all the attributes and consequently the parameters of each event are learned separately. In real-world applications, events appear in most cases dependently. In other words, the occurrence of one event affects the probability of the second occurred event. The multinomial independence assumption hinders capturing the phenomenon of burstiness, *i.e.,* a word that has already appeared in a document, for instance, has a higher probability of appearing again [4, 3]. Thus, many studies have proposed solutions to overcome this deficiency.

Dirichlet multinomial (DM) has been introduced as an alternative to the multinomial distribution [22] for count data modeling. A distribution that combines the multinomial and the Dirichlet as a prior has been successful to tackle the problems of word burstiness, overdispersion and sparsity. In fact, the Dirichlet multinomial model has achieved good results in various applications like analysis of taxonomic abundances in microbiome data [23], novelty detection environment [24], minimizing the cost of computing in cloud [25], and texture modeling [26]. However, the Dirichlet distribution has a restrictive covariance matrix where any two random attributes have to be negatively correlated. In addition, the Dirichlet is limited by the reason of variables with the same mean must have the same covariance which makes the model unrealistic. Taking into account those disadvantages, the author in [27] proposed the use of generalized Dirichlet distribution as an alternative prior to the multinomial distribution. In fact, the generalized Dirichlet distribution [28] allows having equal mean and different values for the covariance to reflect different amounts of prior information. Also, variables can be positively or negatively correlated. Additionally, the generalized Dirichlet has $d + 1$ extra parameters compared to the Dirichlet. The generalized Dirichlet multinomial (GDM) distribution has shown outstanding results in interesting applications namely spatial colour indexing, handwritten digit recognition, text document clustering [27], classification of traffic congestion, detection of unusual events in traffic flows, anomaly detection in crowded scenes, human action recognition [29], and consumption behavior prediction [30].

The Fisher information matrix, introduced by Ronald Fisher in 1922 [31], is one of the most significant measures in information theory in general and a particular Riemannian metric that defines the differential geometric structure. For smooth Riemannian manifolds of probability distributions, the Fisher information is related to the surface area of the associated set which measures the amount of information about an unobserved parameter. The exploitation of this measure appeared in literature in the context of large variety of areas including linear dynamic systems [32], nonlinear models [33], time-series analysis [34], discriminative atom embedding [35], nuclear magnetic resonance [36], and image processing [37]. However, given the fact that computing the exact Fisher information matrix is quite complex and intractable in certain cases, the majority of research works have used asymptotic Fisher information or identity matrix instead [38, 39, 40, 41, 42]. For instance, asymptotic information can be obtained by supposing that all the parameters are independent and there is no correlation between the features of the given data. Besides, in [39], authors assume, for large clusters, that Fisher information of Dirichlet-multinomial can be approximated by to the

Fisher of Dirichlet. However, Paul et al. [43] proved that in practice, clusters sizes are not always large. Thus, they proposed the exact Fisher information matrix for the DM distribution.

To the best of our knowledge, the Fisher information matrix of the generalized Dirichlet multinomial distribution has never been calculated exactly and it was only approximated. For this purpose, we aim to introduce an exact calculation of the Fisher information matrix. The goal of this chapter is to produce an analytically tractable solution to the calculation of Fisher matrix's elements through presenting a parametrization of generalized Dirichlet multinomial mixture model based on the properties of Beta-binomial probability function.

Mixture models have been extensively applied in the last decade for several applications related to computer vision and machine learning [44, 45, 46]. The big challenge in mixture models is learning the model's parameters. One of the standard estimation techniques is the likelihood-based approach which is based on maximizing the log-likelihood function with respect to the mixture model's parameters. Expectation-Maximization [47] is a well-known learning method despite all the deficiencies that may limit the performance such as the initialization and the convergence problem. Different extensions have been proposed [48] as alternatives to resolve these difficulties. The deterministic annealing EM [49] is one of the most successful alternatives obtained by modifying the E-step in which the posterior probability is parametrized through computational temperature and the model's parameters are updated until the initial high-temperature decreases. This extension offers more adaptable results but is still limited as the parameters are updated in the same manner. Another critical issue in mixture models is the choice of the number of components that better describe the data. Various model selection methods have been considered for this challenging aspect such as the minimum description length (MDL), the Akaike's information criterion (AIC), and the minimum message length (MML). One of the most successful model selection methods that have shown good performance for mixture models is based on MML [50].

In addressing the aforementioned problems, we introduce a new approach for count data modeling. We tackle the learning of the generalized Dirichlet multinomial mixture model using a new deterministic-annealing EM approach within a Fisher-scoring algorithm. We improve further the mixture model using MML criterion and the exact Fisher information matrix.

The main contributions of this chapter can be summarized as follows.

(1) We propose an exact calculation of the Fisher information matrix (EFIM) in favour of generalized Dirichlet multinomial (GDM) mixture model

(2) We introduce a new Deterministic-annealing EM learning algorithm for the generalized Dirichlet multinomial based on Fisher scoring algorithm and minimum message length for count data modeling

(3) We cluster different modalities of count data (text, dialogue, image) in several challenging applications namely detecting depression in tweets, dialogue-based emotion recognition, and image-based sentiment analysis as well as evaluating the capability of computing EFIM for the GDM mixture model.

The subsequent sections of this chapter are laid out as follows: Section 2 reviews the basic multinomial-based distributions and the problem of the independence assumption. Then, in Section 3, we present the proposed methodology of computing the EFIM. Section 4 introduces the considered learning method for estimating the model's parameters. Experimental results are illustrated upon training the proposed strategy in Section 5. Lastly, Section 6 concludes the paper as well as propounding future work.

2.2 Background

In this section, we provide in the first place a brief background of Dirichlet multinomial and generalized Dirichlet multinomial. We explain the limitations of DM and highlight some of the GDM advantages to understand the structure of bag-of-scaled-documents that is proposed, initially, for modeling word counts. Following, we point out the poor assumption of independence which is extensively used for simplification but rarely holds [51, 52, 53, 54].

2.2.1 Count data modeling using DM and GDM

The multinomial distribution is a generalization of the Binomial distribution. Generally speaking, when we consider the multinomial distribution, we are modeling the probability of counts of N discrete variables appearing in one of D possible states, where each state of the variables is occurring with the probabilities $P_1, \ldots, P_D$. When we consider the bag-of-words structure, the multinomial distribution is widely used to capture the word frequency information in documents. So, each document is represented by the set of word occurrences and drawn from a multinomial distribution of words.

Let $\vec{X}$ be a vector of counts which follows a multinomial distribution with parameters $\vec{P} = (P_1, \ldots, P_D)$:

$$p(\vec{X}|\vec{P}) = \frac{n!}{\prod_{l=1}^{D} x_l!} \prod_{l=1}^{D} P_l^{x_l} \tag{1}$$

where $n = \sum_{l=1}^{D} x_l$, the parameter P_l is the probability of emitting a word l from the document represented by $\vec{X}$, $0 \leq P_l \leq 1$ and $\sum_l P_l = 1$.

The most popular solution to overcome the limitations and the deficiencies of the multinomial distribution is the Dirichlet multinomial (DM) which is the composition of the Dirichlet and the multinomial [22], and has shown to be competitive with the best known text classification methods by handling the burstiness successfully and accurately. The multinomial Dirichlet distribution is defined over a count vector for each document generated by a multinomial distribution whose parameters are generated by the Dirichlet distribution, called bag-of-scaled-documents model. In this case, the DM is characterized by a joint parameter $\Psi = (\vec{P}, \alpha)$, where α is the actual parameter of interest. Thus, marginalizing out $\vec{P}$ vectors for the multinomial weighted by the Dirichlet distribution gives us the likelihood of a document using the DM approach:

$$\begin{aligned} p(\vec{X}|\boldsymbol{\alpha}) &= \int_{\vec{P}} p(\vec{X}|\vec{P}) p(\vec{P}|\boldsymbol{\alpha}) d\vec{P} \qquad (2) \\ &= \frac{\Gamma(\sum_{l=1}^{D} x_l + 1)\Gamma(\sum_{l=1}^{D} \alpha_l)}{\Gamma(\sum_{l=1}^{D} x_l + \sum_{l=1}^{D} \alpha_l)} \prod_{l=1}^{D} \frac{\Gamma(x_l + \alpha_l)}{\Gamma(\alpha_l)\Gamma(x_l + 1)} \end{aligned}$$

Whereas the Dirichlet distribution is the best-known distribution for being a prior for the multinomial distribution, it has a restrictive negative covariance and constant sum constraint. With the need for completely neutral vectors of proportions, authors in [28] proposed consideration of the generalized Dirichlet distribution. The concept of neutrality is defined as the independence of a vector P_1 from the proportions $P_2/(1 - P_1), \ldots, P_D/(1 - P_1)$. This concept is extended for more than one variable and makes the generalization of the Dirichlet distribution with parameter vector $(\alpha_1, \beta_1, \ldots, \alpha_{D-1}, \beta_{D-1})$ as defined by [28]:

$$p(P_1, ..., P_D) = \prod_{l=1}^{D-1} \frac{\Gamma(\alpha_l + \beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} P_l^{\alpha_l - 1} \Big(1 - \sum_{j=1}^{l} P_j\Big)^{\gamma_l} \tag{3}$$

for $P_1 + P_2 + \cdots + P_D \leq 1$ and $P_l \geq 0$ for $l = 1, \ldots, D$ where $\gamma_l = \beta_l - \alpha_{l+1} - \beta_{l+1}$, $l = 1, \ldots, D-2$, and $\gamma_{D-1} = \beta_{D-1} - 1$.

The generalized Dirichlet distribution can be reduced to a Dirichlet distribution when $\beta_l = \alpha_{l+1} + \beta_{l+1}$ and the univariate case ($l = 2$) corresponds to the Beta distribution.

This distribution arises in various contexts including Bayesian life-testing problems [55], mixture models for pattern recognition [56], and machine learning for image processing [57]. In addition to these utilizations, the generalized Dirichlet was introduced as a prior to the multinomial distribution by [27] for many reasons that make the generalized Dirichlet more flexible than other distributions and especially that is a conjugate to the multinomial distribution.

The joint distribution of a vector $\vec{X}_i = (x_{i1}, \ldots, x_{iD})$ and $\vec{P}$ is defined as the following:

$$p(\vec{X}_i, \vec{P}|\vec{\alpha}, \vec{\beta}) = \frac{\Gamma(\sum_{l=1}^{D} x_{il} + 1)}{\prod_{l=1}^{D} \Gamma(x_{il} + 1)} \prod_{l=1}^{D-1} \frac{\Gamma(\alpha_l + \beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} P_l^{\alpha'_l - 1} \Big(1 - \sum_{j=1}^{l} P_j\Big)^{\gamma'_l} \tag{4}$$

where $\alpha'_l = \alpha_l + x_{il}$ and $\beta'_l = \beta_l + x_{il+1} + ... + x_{iD}$ for $l = 1, \ldots, D-1$, $\gamma'_l = \beta'_l - \alpha'_{l+1} - \beta'_{l+1}$ for $l = 1, \ldots, D-2$ and $\gamma'_{D-1} = \beta'_{D-1} - 1$.

Marginalizing out $\vec{P}$, we obtain the generalized Dirichlet multinomial (GDM) distribution of $\vec{X}_i$ with parameters $(\alpha'_1, \beta'_1, \ldots, \alpha'_{D-1}, \beta'_{D-1})$ as follows:

$$\begin{aligned} p(\vec{X}_i|\vec{\alpha}', \vec{\beta}') &= \int_{\vec{P}} p(\vec{X}_i, \vec{P}|\vec{\alpha}, \vec{\beta}) d\vec{P} \\ &= \frac{\Gamma(\sum_{l=1}^{D} x_{il} + 1)}{\prod_{l=1}^{D} \Gamma(x_{il} + 1)} \prod_{l=1}^{D-1} \frac{\Gamma(\alpha_l + \beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} \int_{\vec{P}} P_l^{\alpha'_l - 1} \Big(1 - \sum_{j=1}^{l} P_j\Big)^{\gamma'_l} d\vec{P} \\ &= \frac{\Gamma(\sum_{l=1}^{D} x_{il} + 1)}{\prod_{l=1}^{D} \Gamma(x_{il} + 1)} \prod_{l=1}^{D-1} \frac{\Gamma(\alpha_l + \beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} \prod_{l=1}^{D-1} \frac{\Gamma(\alpha'_l)\Gamma(\beta'_l)}{\Gamma(\alpha'_l + \beta'_l)} \end{aligned} \tag{5}$$

2.2.2 Independence assumption

In this section, we inaugurate with Naïve Bayes assumption considered extensively in text retrieval and classification. Then, we put forward the independence assumption applied by Fisher information matrix.

In information retrieval, in a bag-of-words structure, the Naïve Bayes assumption assumes that words in a document are independent. Suppose we have a collection of documents $\mathcal{X} = (\vec{X}_1, \dots, \vec{X}_N)$, and each document $\vec{X}_i = (x_{i1}, \dots, x_{iD})$ is associated with features represented by D words. If we make the Naïve Bayes assumption, then the conditional distribution of the document $\vec{X}_i$ given a class c_k is defined by:

$$p(\vec{X}_i|c_k) = \prod_{l=1}^{D} p(x_{il}|c_k) \tag{6}$$

We interpret this equation as all features are independent in a document of class c_k. We also infer that each $\vec{X}_i$ is independent of any other $\vec{X}_i'$ in the same document. This results in simpler, faster, and easier way to implement models where $p(x_{il}|c_k)$ is employed with relatively fewer parameters, but believing the data to be independent, when in fact, is not. This entails an inappropriate balancing for one class over another in regard to skewed data. Besides, the Naïve Bayes assumption gives rise to bias in the weights of feature vectors.

The independence assumption in the Fisher information matrix is to consider only the diagonal elements or to address instead the identity matrix. The Fisher information reflects how a model is sensitive, so, ignoring this information to be an identity matrix seems to imply that the amount of information in such parameters of the considered model is equally distributed and unitary. For simplicity reasons, this assumption is assumed, but may not be a good approximation over all types of data and in particular count data. In this context, for the several advantages early mentioned of the GDM, we propose the calculation of the exact Fisher information matrix for GDM.

2.3 Derivation of the Exact Fisher Information Matrix for GDM

2.3.1 Parametrization of GDM mixture model

The parametrization process is a transformation that acts only on the parameters of the distribution. We consider the generalized Dirichlet multinomial distribution where the parameters of interest are $\vec{\alpha}$ and $\vec{\beta}$. We display the probability density function of GDM in terms of factorial and Dirichlet integral:

$$p(\vec{X}_i|\vec{\alpha},\vec{\beta}) \;=\; \frac{\sum_{l=1}^{D} x_{il}!}{\prod_{l=1}^{D} x_{il}!} \prod_{l=1}^{D-1} \frac{\Gamma(\alpha_l+\beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} \int_{\vec{P}} P_l^{\alpha'_l-1}\Big(1-\sum_{j=1}^{l} P_j\Big)^{\gamma'_l} d\vec{P} \tag{7}$$

Noting $m = \sum_{l=1}^{D} x_{il}$, and considering the variables $\alpha'_l = \alpha_l + x_{il}$ and $\beta'_l = \beta_l + \sum_{k=l+1}^{D} x_{ik}$, the GDM can be written also as follows:

$$p(\vec{X}_i|\vec{\alpha},\vec{\beta}) \;=\; \binom{m}{x_{i1}\dots x_{iD}} \prod_{l=1}^{D-1} \frac{\Gamma(\alpha_l+x_{il})}{\Gamma(\alpha_l)} \frac{\Gamma(\beta_l+\sum_{k=l+1}^{D} x_{ik})}{\Gamma(\beta_l)} \frac{\Gamma(\alpha_l+\beta_l)}{\Gamma(\alpha_l+\beta_l+\sum_{k=l}^{D} x_{ik})} \tag{8}$$

We consider the following new parameters that enable to profit from the properties of the Beta-binomial probability function:

$$\Theta_l = \frac{1}{\alpha_l+\beta_l} \tag{9}$$

$$\pi_l = \frac{\alpha_l}{\alpha_l+\beta_l} \tag{10}$$

Further, given that $\Gamma(X) = (X-1)!$, we obtain:

$$p(\vec{X}_i|\Phi) \;=\; \binom{m}{x_{i1}\dots x_{iD}} \prod_{l=1}^{D-1} \frac{\pi_l \dots [\pi_l+(x_{il}-1)\Theta_l](1-\pi_l)\dots[1-\pi_l+(y_{il+1}-1)\Theta_l]}{1\dots[1+(y_{il}-1)\Theta_l]} \tag{11}$$

where, $y_{il} = \sum_{k=l}^{D} x_{ik},\ y_{il} = y_{il+1} + x_{il}$

$$p(\vec{X}_i|\Phi) \quad = \quad \binom{m}{x_{i1}\dots x_{iD}} \prod_{l=1}^{D-1} \frac{\prod_{r=1}^{x_{il}}[\pi_l + (r-1)\Theta_l]}{\prod_{r=1}^{y_{il}}[1+(r-1)\Theta_l]} \frac{\prod_{r=1}^{y_{il+1}}[1-\pi_l+(r-1)\Theta_l]}{\prod_{r=1}^{y_{il}}[1+(r-1)\Theta_l]} \tag{12}$$

and Φ is the new set of parameters

$$\Phi = (\Theta_1, \dots, \Theta_{D-1}, \pi_1, \dots, \pi_{D-1})$$

A mixture of generalized Dirichlet multinomial with M components is given by:

$$p(\vec{X}_i|\mathbf{\Phi}) = \sum_{j=1}^{M} p_j p(\vec{X}_i|\Phi_j) \tag{13}$$

where $\mathbf{\Phi} = (\pi_{11}, \dots, \pi_{MD-1}, \Theta_{11}, \dots, \Theta_{MD-1}, p_1, \dots, p_M)$, p_j $(0 < p_j \leq 1$ and $\sum_{j=1}^{M} p_j = 1)$ are the mixing weight, and $p(\vec{X}_i|\Phi_j)$ is a GDM distribution.

2.3.2 Derivation of the EFIM

Given a set of independent vectors $\mathcal{X} = (\vec{X}_1, \dots, \vec{X}_N)$, the complete log-likelihood corresponding to the generalized Dirichlet multinomial mixture model, apart from a constant is:

$$\begin{aligned}\log L(\mathbf{\Phi}, \mathcal{X}) \quad &\propto \quad \sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{X}_i)\log(p(j)p(\vec{X}|\Phi_j)) \\ &\propto \quad \sum_{i=1}^{N}\sum_{j=1}^{M} p(j|\vec{X}_i)\Big\{\log(p(j)) + \sum_{l=1}^{D-1}\Big[\sum_{r=1}^{x_{il}}\log(\pi_{jl} + (r-1)\Theta_{jl}) \\ &+ \quad \sum_{r=1}^{y_{il+1}}\log(1-\pi_{jl}+(r-1)\Theta_{jl}) - \sum_{r=1}^{y_{il}}\log(1+(r-1)\Theta_{jl})\Big]\Big\}\end{aligned}$$

where

$$p(j|\vec{X}_i) = \frac{p_j p(\vec{X}_i|\Phi_j)}{\sum_{j=1}^{M} p_j p(\vec{X}_i|\Phi_j)} \tag{14}$$

represents the posterior probability that a vector $\vec{X}_i$ is affected to the component j of the mixture.

The exact Fisher information matrix, for a multidimensional parameter space Φ, takes the following form:

$$F = E(-H(\Phi)) \tag{15}$$

where H is the second derivative matrix known as the Hessian matrix and below, the Fisher information is specified for each dimension of the parameters in the considered mixture components:

$$F = \begin{bmatrix} F_{11} & F_{12} & \dots & \dots & F_{1D-1} \\ F_{21} & F_{22} & \dots & \dots & F_{2D-1} \\ \dots & \dots & \dots & \dots & \dots \\ F_{j1} & F_{j2} & \dots & \dots & F_{jD-1} \\ \dots & \dots & \dots & \dots & \dots \\ F_{M1} & F_{M2} & \dots & \dots & F_{MD-1} \end{bmatrix} \tag{16}$$

where for $j = 1, \dots, M$ and $l = 1, \dots, D-1$, the F_{jl} is defined as the expected shape of the likelihood function in the parameter space where the diagonal elements are the variance of parameters π and Θ and the off-diagonal denotes the correlation between the inferred parameters according to the following formula:

$$F_{jl} = \begin{bmatrix} E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial \pi_{jl}^2}\Big] & E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial \pi_{jl}\Theta_{jl}}\Big] \\ \\ E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial \Theta_{jl}\pi_{jl}}\Big] & E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial \Theta_{jl}^2}\Big] \end{bmatrix} \tag{17}$$

We start, here, by calculating the expectation of the first element of the Fisher information matrix:

$$\begin{aligned} E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial \pi_{jl}^2}\Big] &= E\Big[\sum_{i=1}^{N} p(j|\vec{X}_i)\Big(\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2} \\ &+ \sum_{r=1}^{y_{il}+1} \frac{-1}{\{1-\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big)\Big] \end{aligned} \tag{18}$$

Taking into account the linearity properties of the expectation of discrete numbers, we get:

$$
\begin{aligned}
E\Big[\frac{-\partial^2 \log L(\boldsymbol{\Phi}, \mathcal{X})}{\partial \pi_{jl}^2}\Big] &= \sum_{i=1}^{N} p(j|\vec{X}_i) E\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big] \\
&+ E\Big[\sum_{r=1}^{y_{il}+1} \frac{-1}{\{1 - \pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big]
\end{aligned} \tag{19}
$$

Knowing that the random variable $\vec{X}_i$ follows the generalized Dirichlet multinomial distribution. Then, the expectation of the observation $\vec{X}_i = (x_{i1}, \ldots, x_{iD})$ for the first term of the above equation is defined as:

$$
E\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big] = \sum_{x_{i1}, \ldots, x_{iD}}^{m} p(X_i) \sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2} \tag{20}
$$

where $m = \sum_{l=1}^{D} x_{il}$ and in a simplified manner, we can write down the above equation for the k-th element only:

$$
\begin{aligned}
E\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big] &= \sum_{x_{il}=0}^{m} \sum_{x_{i1}, \ldots, x_{il-1}, x_{il+1}, \ldots, x_{iD}}^{m - x_{il}} p(x_{i1}, \ldots, x_{il}, \ldots, x_{iD}) \\
&\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big]
\end{aligned} \tag{21}
$$

Now, if we take $x_{il} = 1$,

$$
\begin{aligned}
E\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big] &= \sum_{x_{i1}, \ldots, x_{il-1}, x_{il+1}, \ldots, x_{iD}}^{m-1} p(x_{i1}, \ldots, 1, \ldots, x_{iD}) \\
&\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}\Big] \\
&= \frac{p(1)}{\{\pi_{jl} + (r-1)\Theta_{jl}\}^2}
\end{aligned} \tag{22}
$$

where $p(1) = p(x_{il})$ is the Beta-binomial probability function. Through this result, we can generalize our computation of the expectation to come by:

$$\begin{aligned} E\Big[\sum_{r=1}^{x_{il}} \frac{1}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2}\Big] &= \sum_{x_{il}=0}^{m}\sum_{r=1}^{x_{il}} \frac{p(x_{il})}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} \\ &= \sum_{r=1}^{m} \frac{p(x_{il}\geq r)}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} \end{aligned} \tag{23}$$

By the same token, the other Fisher information elements are determined. We supply in the following the second and mixed derivatives of the parameters:

$$\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial\Theta_{jl}\pi_{jl}} = \Big[\sum_{r=1}^{x_{il}} \frac{r-1}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} - \sum_{r=1}^{y_{il}+1} \frac{r-1}{\{1-\pi_{jl}+(r-1)\Theta_{jl}\}^2}\Big]$$

$$\begin{aligned} \frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial\Theta_{jl}^2} &= \Big[\sum_{r=1}^{x_{il}} \frac{(r-1)^2}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} + \sum_{r=1}^{y_{il}+1} \frac{(r-1)^2}{\{1-\pi_{jl}+(r-1)\Theta_{jl}\}^2} \\ &- \sum_{r=1}^{y_{il}} \frac{(r-1)^2}{\{1+(r-1)\Theta_{jl}\}^2}\Big] \end{aligned} \tag{24}$$

By calculating the expectation of the second derivative of $\log L(\mathbf{\Phi},\mathcal{X})$ with respect to the set of parameters, we obtain the elements of the Fisher information as follows:

$$\begin{aligned} E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial\pi_{jl}^2}\Big] &= \sum_{i=1}^{N} p(j|\vec{X}_i)\Big|\sum_{r=1}^{m} \frac{P(x_{il}\geq r)}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} \\ &- \frac{P(y_{il+1}\geq r)}{\{(1-\pi_{jl})+(r-1)\Theta_{jl}\}^2}\Big| \end{aligned} \tag{25}$$

$$\begin{aligned} E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial\Theta_{jl}\pi_{jl}}\Big] &= \sum_{i=1}^{N} p(j|\vec{X}_i)\Big|\sum_{r=1}^{m}(r-1) \frac{P(x_{il}\geq r)}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} \\ &- \frac{P(y_{il+1}\geq r)}{\{(1-\pi_{jl})+(r-1)\Theta_{jl}\}^2}\Big| \end{aligned} \tag{26}$$

$$E\Big[\frac{-\partial^2 \log L(\mathbf{\Phi},\mathcal{X})}{\partial\Theta_{jl}^2}\Big] = \sum_{i=1}^{N} p(j|\vec{X}_i)\Big|\sum_{r=1}^{m}(r-1)^2\frac{P(x_{il}\geq r)}{\{\pi_{jl}+(r-1)\Theta_{jl}\}^2} + \frac{P(y_{il+1}\geq r)}{\{(1-\pi_{jl})+(r-1)\Theta_{jl}\}^2} - \frac{P(y_{il}\geq r)}{\{1+(r-1)\Theta_{jl}\}^2}\Big| \tag{27}$$

where $p(x_{il} \geq r)$ has a Beta-binomial probability density function of the vector x_{il} that should be bigger than the sum of the features.

2.4 Generalized Dirichlet Multinomial Mixture Estimation and Selection

2.4.1 Fisher Scoring estimation algorithm

The Fisher scoring algorithm is a maximum likelihood estimation method that requires the inverse of the Fisher information matrix in order to converge to a local maximum [58]. The estimation of the considered parameters using Fisher scoring algorithm at iteration $(t+1)$ are:

$$\Phi_{jl}^{(t+1)} = \Phi_{jl}^{t} + [F_{jl}(\Phi)]^{-1}S_{jl}(\Phi^t) \tag{28}$$

where Φ_{jl}^{t} is the vector of the estimated parameters at iteration t, $S_{jl}(\Phi^t)$ is the score vector (gradient of the log-likelihood) at iteration t and $F_{jl}(\Phi)$ is jl element of the Fisher information matrix for the parameter space Φ.

By computing the gradient of $\log L(\mathbf{\Phi},\mathcal{X})$ with respect to the model parameters π_{jl},Θ_{jl}, we obtain:

$$\begin{aligned} S_{jl}(\pi) &= \nabla_{\pi_{jl}} \log L(\mathbf{\Phi},\mathcal{X}) \\ &= \sum_{i=1}^{N} p(j|\vec{X}_i)\Big|\sum_{r=1}^{x_{il}}\frac{1}{\pi_{jl}+(r-1)\Theta_{jl}} + \sum_{r=1}^{y_{il+1}}\frac{-1}{(1-\pi_{jl})+(r-1)\Theta_{jl}}\Big| \end{aligned}$$

$$\begin{aligned} S_{jl}(\Theta) &= \nabla_{\Theta_{jl}} \log L(\mathbf{\Phi}, \mathcal{X}) \\ &= \sum_{i=1}^{N} p(j|\vec{X}_i) \Big| \sum_{r=1}^{x_{il}} \frac{r-1}{\pi_{jl} + (r-1)\Theta_{jl}} + \sum_{r=1}^{y_{il+1}} \frac{r-1}{(1-\pi_{jl}) + (r-1)\Theta_{jl}} \\ &- \sum_{r=1}^{y_{il}} \frac{r-1}{1 + (r-1)\Theta_{jl}} \Big| \end{aligned} \tag{29}$$

$j = 1, ..., M,\ \ l = 1, \ldots, D-1.$

2.4.2 Minimum Message Length

The minimum message length (MML) [50] selection criterion for mixture models has shown to be a good choice to select the number of components in numerous applications by avoiding overfitting and underfitting problems when modeling data. MML is based on Bayesian information theory, where it uses explicitly the prior distribution of the parameters and the Fisher information matrix. Using MML, the optimal number of components in the mixture is obtained by minimizing the following function:

$$MML(\mathbf{\Phi}, \mathcal{X}) = -\log(p(\Phi)) - L(\mathbf{\Phi}, \mathcal{X}) + \frac{1}{2}\log|F(\Phi)| + \frac{N_p}{2} + \frac{N_p}{2}\log K_{N_p} \tag{30}$$

where $p(\mathbf{\Phi})$ is the prior probability, $L(\mathbf{\Phi}, \mathcal{X})$ is the complete likelihood, $|F(\Phi)|$ is the determinant of the exact Fisher information matrix, N_p is the number of parameters to be estimated and is equal to $M(2D+1)$ in our case, and K_{N_p} is a lattice constant.

The key of this selection method is the adequate choice of the prior parameter $p(\mathbf{\Phi})$ where a possible selection for Θ and π parameters is the Gamma distribution and for the mixing weight, we consider a uniform prior.

$$p(\mathbf{\Phi}) = (M-1)! \prod_{j=1}^{M} \prod_{l=1}^{D-1} p_{Gamma}(\pi_{jl})\ \ p_{Gamma}(\Theta_{jl}) \tag{31}$$

And we consider the exact Fisher information matrix in MML instead of considering an approximated matrix:

$$|F(\Phi)| = |F(p_1, \ldots, p_M)| \prod_{j=1}^{M} \prod_{l=1}^{D-1} |F_{jl}| \tag{32}$$

where the mixing parameters $p_1, \ldots, p_M$ can be considered as the parameters of a multinomial distribution for which the determinant of the Fisher information $|F(p_1, \ldots, p_M)|$ is [50]:

$$|F(p_1, \ldots, p_M)| = \frac{N}{\prod_{j=1}^{M} p_j} \tag{33}$$

2.4.3 Convergence and initialization issues

The initialization and convergence of the likelihood function have a considerable influence on the entire algorithm to estimate the appropriate parameters. Initialization, per se, influences the determination of GDM parameters in the first place. Considering, for instance, the method of moments as for the generalized Dirichlet distribution. Using the sample estimate of the first and the second moments, the calculation of α and β are as follows:
for $l = 1, \ldots, D-1$

$$\alpha_l = \frac{a_l\ \mu_l\ B_{l-1} - \mu_l\ (S_{l,l} + \mu_l^2)}{A_{l-1}\ (S_{l,l} + \mu_l^2) - a_l\ \mu_l\ B_{l-1}} \tag{34}$$

$$\beta_l = \frac{\alpha_l (A_{l-1} - \mu_l)}{\mu_l} \tag{35}$$

where, $A_0 = B_0 = 1$ and for $l = 1, \ldots, D-1$

$$A_l = \prod_{k=1}^{l} \frac{\beta_k}{\alpha_k + \beta_k} \tag{36}$$

$$B_l = \prod_{k=1}^{l} \frac{\beta_k(\beta_k + 1)}{(\alpha_k + \beta_k)(\alpha_k + \beta_k + 1)} \tag{37}$$

$$a_l = \frac{\alpha_l}{\alpha_l + \beta_l} \tag{38}$$

and μ_l and $S_{l,l}$ are the mean and the variance of the generalized Dirichlet given by [28]:

$$E(P_l) = \frac{\alpha_l}{\alpha_l + \beta_l} \prod_{k=1}^{l-1} \frac{\beta_k}{\alpha_k + \beta_k} \tag{39}$$

$$Var(P_l) = E(P_l)\Big(\frac{\alpha_l + 1}{\alpha_l + \beta_l + 1} \prod_{k=1}^{l-1} \frac{\beta_k + 1}{\alpha_k + \beta_k + 1} - E(P_l)\Big) \tag{40}$$

$l = 1, \ldots, D-1$, and $Var(P_l)$ is simplified by:

$$S_{l,l} = Var(P_l) = \frac{\alpha_l + 1}{\alpha_l + \beta_l + 1} a_l B_{l-1} - \mu_l^2 \tag{41}$$

As it involves the parameters of the generalized Dirichlet and we are considering a parametrization of the GDM, the initial parameters for the proposed model are obtained through the following:

$$\pi_{l_0} = \frac{\alpha_l}{\alpha_l + \beta_l} \tag{42}$$

$$\Theta_{l_0} = \frac{1}{\alpha_l + \beta_l} \tag{43}$$

$l_0 = 1, \ldots, D-1$,

Given the initialization step and the proposed Fisher scoring estimation algorithm, the whole learning approach so-called $\mathcal{FSGDM}$ is updated through an iterative principle of Deterministic annealing Expectation-Maximization [47] algorithm. Thus, in the E-step, the posterior probability is parametrized by a temperature parameter τ, and, in M-step, the parameters are updated through Fisher scoring algorithm.

$$p(j|\vec{X}_i) = \frac{\Big(p_j p(\vec{X}_i|\Phi_j)\Big)^\tau}{\sum_{j=1}^{M} \Big(p_j p(\vec{X}_i|\Phi_j)\Big)^\tau} \tag{44}$$

To test the convergence, we set a small threshold between each two-consecutive log-likelihood ($\approx 10^{-3}$) at which the algorithm should stop (Algorithm 1). Experimentally, the initial parameter τ_{min} is set to a small value (0.2) and we choose as temperature scheduling the following constant $const = 5$.

Algorithm 1: $\mathcal{FSGDM}$ learning algorithm

```
Input: Dataset $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$;
Output: Parameters $\vec{\Phi}^*$, number of components $M^*$;
foreach Number of components $M$ do
    Cluster the data $\mathcal{X}$ using K-means;
    Initialize $\alpha$ and $\beta$ using Method of Moments through Eq. 34 and 35;
    Reparametrize the initial parameters $\pi_{l_0}$ and $\Theta_{l_0}$ $(l_0 = 1, \ldots, D-1)$ using Eq. 42, 43;
    Initialize the temperature parameter $\tau = \tau_{min} << 1$;
    while $\tau \leq 1$ do
        repeat
            foreach Component $j$ do
                E-step:;
                Estimate the posterior distribution $p(j|\vec{X}_i)$ $(i = 1, \ldots, N)$ using Eq. 44;
                M-step:;
                Estimate the mixing weight components using $p_j = \frac{1}{N}\sum_{i=1}^{N} p(j|\vec{X}_i)$;
                Update the parameters $\pi_{jl}$ and $\Theta_{jl}$ $(l = 1, \ldots, D-1)$ using Fisher scoring
                  algorithm Eq. 28;
            end
        until Convergence of Log-Likelihood;
        Update the temperature parameter $\tau = \tau \times const$;
    end
    Update the MML criterion using Eq. 30;
end
Select the optimal $M^*$ such that: $M^* = \underset{M}{argmin} MessLength(M)$;
```

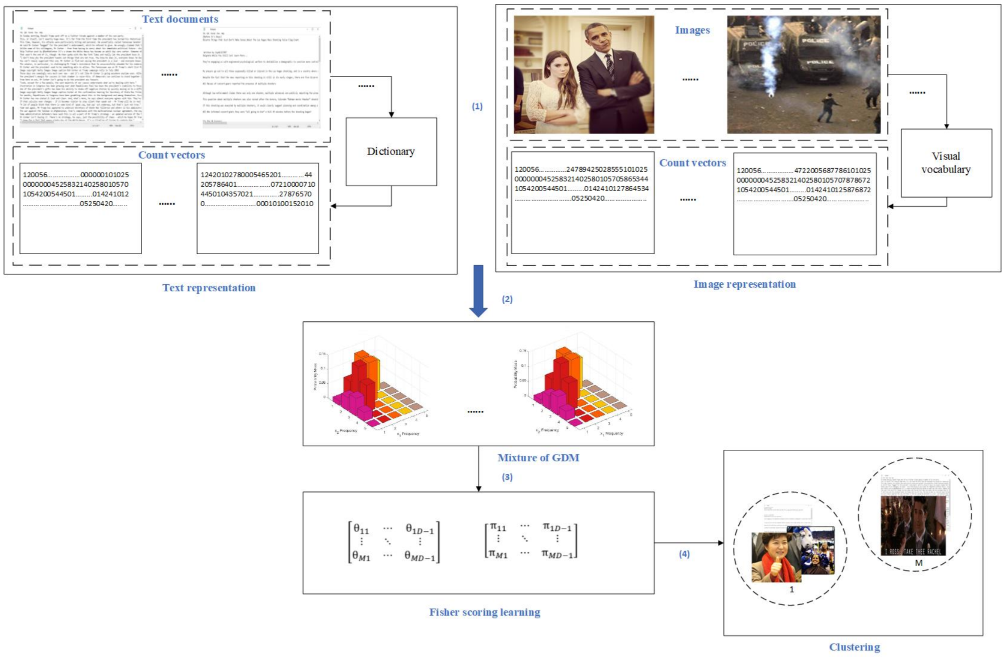


Figure 2.1: Adopted $\mathcal{FSGDM}$ model strategy for experimental analysis

2.5 Experimental analysis

In this chapter, we validate our proposed model on three distinctive applications: detecting depression in tweets, dialogue-based emotion recognition, and image-based sentiment analysis. We use texts and images datasets to confirm the robustness of our approach. The illustration of the proposed framework is displayed in Figure 2.1.

2.5.1 Detecting depression in tweets

Depression is one of the mental health disorders, as mentioned in the World Health Organization's Comprehensive Mental Health Action Plan 2013-2020 [59], more than 300 million people are affected by depression worldwide. Currently, most people use social media to express their emotional states, happy moments, anxiety, and even sadness which allows recognizing depressive users and to reveal earlier serious problems such as suicide. Previously, machine learning techniques have attempted to automate depression detection with the need of psychological experts to assist in intelligent mental-health support. To catch people especially teenagers in an early stage of suffering, tweets have been employed to detect their attitudes and behaviours. In our experiments, we choose a tweet-text dataset [1] which is a combination of the Sentiment140 dataset [60] and depressive tweets generated from the TWINT tool [2]. Sentiment140 dataset contains three polarity tweets (positive, negative, neutral), only the positive tweets are extracted from the mentioned data (8000 tweets) and depressive tweets (2314 tweets) from the TWINT tool to have a total of 10314 tweets. We generate a Bag-of-words from the total number of tweets, Figure 2.2 displays the bag-of-positive words and the bag-of-depressive words. We remove as a preprocessing step all the stop, short, and rare words from the generated vocabulary. Along with this, each tweet is defined by a vector of counts containing the number of occurrences for each given word from the vocabulary.

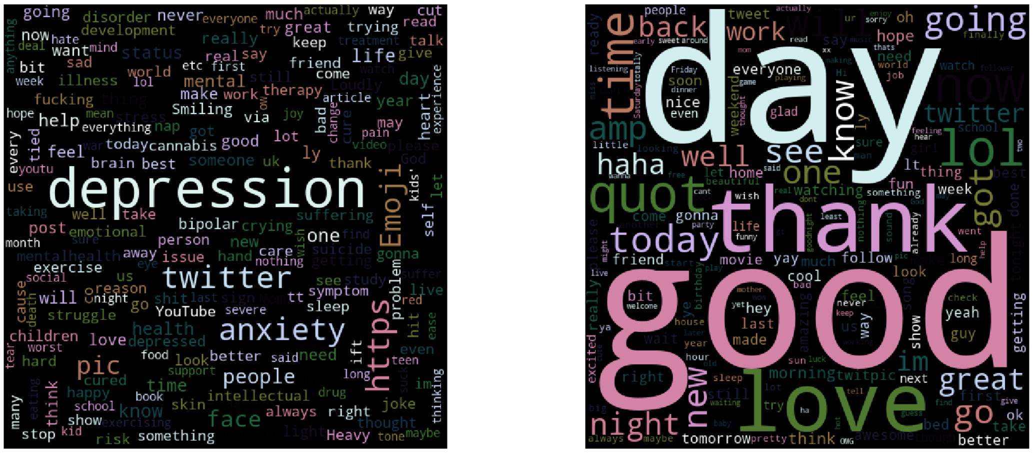
We evaluate our results using accuracy, precision, recall, and F-measure metrics. We compare the novel $\mathcal{FSGDM}$ using the exact Fisher information matrix with related multinomial models: Dirichlet-multinomial using Expectation-Maximization (EM) algorithm, Deterministic annealing EM, Fisher-scoring learning method, and Generalized Dirichlet-multinomial with EM and DAEM inference. Table 2.1 indicates that $\mathcal{FSGDM}$ algorithm is able to recognize depressed Twitter users with 88.79% accuracy. $\mathcal{FSGDM}$ outperforms the other related multinomial models due to the strong characteristics of the Fisher information which consider the dependence of the different features in text documents. Besides, the results of the Fisher-scoring prove the highest difference between the other learning method (EM, DAEM) with more than 10% increase in the precision, recall, and F-measure. Further, we compare the performance of the proposed clustering algorithm with classification methods such as SVM, Logistic regression, and Bayes Theorem. It is clear in Table 2.1 how the results of $\mathcal{FSGDM}$ are comparable with SVM despite that our approach is completely unsupervised which proves more the robustness of the proposed algorithm.

2.5.2 Dialogue-based emotion recognition

Emotion analysis is the process of mining text data to identify and classify subjective information basically into positive, negative, or neutral emotions which help to understand sentiments in social media. This process has tremendously dispersed to include opinion mining, client reviews,

[1] https://github.com/viritaromero/Detecting-Depression-in-Tweets

[2] https://github.com/twintproject/twint



(a) Depressive words

(b) Positive words

Figure 2.2: Bag-of-words for Depression dataset

Table 2.1: Evaluation results for Depression dataset

Models	Accuracy	Precision	Recall	F-measure
Classification methods				
Bayes theorem	79.67	92.85	26.00	40.00
Logistic regression	99.56	99.50	98.00	98.00
SVM	99.22	99.00	98.00	98.00
Clustering methods				
DM+EM	65.37	50.89	50.87	50.88
DM+DAEM	66.27	52.47	52.38	52.42
DM+FS	75.64	42.77	67.29	68.39
GDM+EM	76.58	60.25	64.20	61.30
GDM+DAEM	79.44	76.42	71.91	73.41
$\mathcal{FSGDM}$	88.79	89.35	88.80	85.39

and more which becomes of great interest for several applications such as marketing, business analysis, election, and tourism. Therefore, significant efforts have focused on extracting and gathering information from social media, review websites, blogs, and forums. Emotion recognition has been handled initially at the document level, after, at the sentence level, and recently at the sub-sentence level. Major parts of techniques and methods in solving the problem of sentiment analysis have been done on a document level. In our work, we are interested in different types of data extracted from the famous TV shows “Friends”. Indeed, we evaluate our model $\mathcal{FSGDM}$ on a new challenging

Table 2.2: Description of EmotionLines dataset

Dataset	EmotionLines
Number of utterances	29,245
Number of dialogues	2,000
Speakers	Joey, Monica, Rachel, Phoebe, Chandler, Ross
Emotions	neutral, non-neutral, joy, fear, surprise, disgust, anger, sadness

dataset namely the EmotionLines dataset [3][61] that consists of seven emotions (non-neutral, joy, fear, surprise, disgust, anger, sadness) extra the neutral emotion. This dataset is composed of $2,000$ dialogues and each dialogue is described by speaker, utterance, emotion, and annotation which results in a total of 29,245 utterances (Table 2.2). For our experiments, we perform two scenarios, one at the sentence level, for each dialogue, we predict the emotion concerning the labeled sentence. The second one, at the dialogue level, we extract all the utterances from the totality of 2,000 dialogues, and we construct a vocabulary using the bag-of-words approach. Thus, each utterance is represented as a vector of counts.

At the sentence level, we model each dialogue apart. We construct the vocabulary W and, then, each utterance of T words is encoded as a W-dimensional feature vector. The number of emotions in one dialogue is different to another. For example, in Table 2.3, the feelings of speakers are non-neutral, neutral, and fear, but, in Table 2.4, the sentiments are joy, neutral, and fear. For each sentence, the emotion depends on the context, where within a dialogue, there is a sense of inter-dependency with a link to the speaker's emotions. For that, we can see the benefits of using the exact Fisher information matrix when considering the dependence between the feature vectors is significant. Hence, in Table 2.3, the recognition rate is 71.42% when only two predicted labels are erroneous. As for the second example in Table 2.4, the neutral emotion was wrongfully predicted as joy, and the fear sentiment of the sentence "Uh-Oh!" as neutral. Even for the challenging data of this example, the proposed model $\mathcal{FSGDM}$ achieves an 80% recognition rate.

At the dialogue level, we evaluate our model $\mathcal{FSGDM}$ with others multinomial-based models and the algorithms proposed in [61]. We employed the same metric (weighted and unweighted accuracy) defined as follows:

[3] http://doraemon.iis.sinica.edu.tw/emotionlines/index.html

$$WA = \sum_{j \in C} w_j a_j \tag{45}$$

$$UWA = \frac{1}{|C|} \sum_{j \in C} a_j \tag{46}$$

where w_j is the percentage of utterances in emotion class j and a_j denotes the accuracy of emotion in class j.

Table 2.5 displays the results of $\mathcal{FSGDM}$ where the weighted accuracy and the unweighted accuracy percentages were increased by $\approx 4\%$ as compared to deep learning algorithms (CNN, CNN-BiLSTM) and by $\approx 20\%$ relative to the count data modeling techniques. This proves again the importance of considering the correlation of the features and the high performance of the proposed framework, which owes its success to the Riemannian manifold properties, that is robust to such challenging dataset and the unbalancing nature of the emotion label distribution.

Table 2.3: Predicting Emotions from Friends TV scripts using $\mathcal{FSGDM}$ for a single dialogue between Rachel and Phoebe

Speaker	Dialogue	Emotion	Predicted emotion
Rachel	Well Phoebe, we gotta do something!	Non-neutral	Non-neutral
Rachel	Well, you know.	Neutral	Fear
Rachel	I mean there's no way Joey's gonna make it in time.	Fear	Fear
Rachel	So, I'm gonna go through the hotel and see if there's any other weddings going on.	Neutral	Non-neutral
Phoebe	Okay. Oh, but don't tell them Monica is pregnant because they frown on that.	Neutral	Neutral
Rachel	Okay	Neutral	Neutral

2.5.3 Image-based sentiment analysis

In this section, we approach the sentiment analysis from visual content. Human emotions pose a quiet set of challenges that text data are not enough to understand the expressing issues of human beings. Concerned about the non-verbal communication, there is room for visual information that

Table 2.4: Predicting Emotions from Friends TV scripts using $\mathcal{FSGDM}$ for a single dialogue between Joey and Monica

Speaker	Dialogue	Emotion	Predicted emotion
Joey	Hey Monica, it's Joey!	Joy	Joy
Monica	Hey Joey! Aww, you remembered even though you're a big star!	Joy	Joy
Joey	Aw, come on! It'll be years before I forget you!	Neutral	Joy
Monica	Joey, what's it like on a movie set, huh?	Neutral	Neutral
Monica	Do you have a dressing room?	Neutral	Neutral
Monica	Do you have a chair with your name on it?	Neutral	Neutral
Joey	Uh, well yeah-yeah, I've got all of that going on.	Neutral	Neutral
Joey	Yeah, listen uh, I want you to make sure you tell Chandler that he couldn't have been more wrong!	Neutral	Neutral
Joey	Uh-oh!	Fear	Neutral
Joey	Everybody smile! Okay, thanks a lot! Enjoy your stay	Joy	Joy

provides more thoughts about people's opinions and sentiments. So far, limited efforts have focused on sentiment analysis from visual content for instance images and videos. The closest that comes to sentiment analysis from visual content are concerning facial expression and user intent which are limited to low-level features. Here, we are interested in visual sentiment analysis where we take advantage of SentiBank dataset [62] created for the purpose of mid-level concept representation of images. This dataset contains 0,5 million images crawled from social media, YouTube and Flickr containing more than 3,000 ANP (Adjective Noun Pairs). Accordingly, the SentiBank consists of image-text combined from tweets where the overall is illustrated into positive and negative sentiments, see Figure 2.3. The images in SentiBank dataset are arranged by adopting the bag-of-visual-words approach. We extract in the first step the visual features using SIFT descriptors for a patch of 16×16 pixels computed over an 8-pixel spacing grid. After, we cluster the obtained visual features to build the vocabulary using K-means algorithm. Thus, each image is defined by a vector of frequencies of visual words.

Table 2.5: Weighted and unweighted accuracy on Friends TV show dataset

	Weighted Accuracy	Unweighted Accuracy
CNN [61]	59.2	45.2
CNN-BiLSTM [61]	63.9	43.1
DM+EM	33.71	33.76
DM+DAEM	34.40	33.97
DM+FS	36.99	37.01
GDM+EM	46.56	42.08
GDM+DAEM	46.66	42.10
$\mathcal{FSGDM}$	66.09	46.60

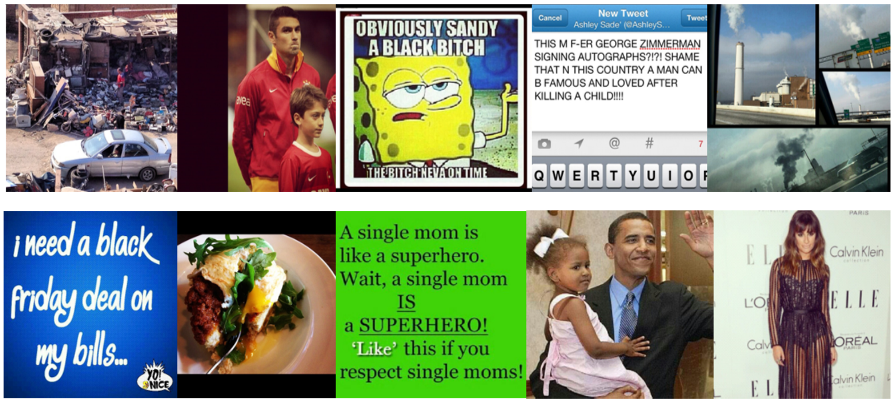

OBVIOUSLY SANDY A BLACK BITCH
THE BITCH NEVA ON TIME
Cancel
New Tweet
Tweet
THIS M F-ER GEORGE ZIMMERMAN SIGNING AUTOGRAPHS?!?! SHAME THAT N THIS COUNTRY A MAN CAN B FAMOUS AND LOVED AFTER KILLING A CHILD!!!!
Q W E R T Y U I O P
i need a black friday deal on my bills...
YO! NICE
A single mom is like a superhero. Wait, a single mom IS a SUPERHERO! 'Like' this if you respect single moms!
ELLE
Calvin Klein
L'OREAL

Figure 2.3: Samples images from SENTIBANK dataset. First row presents samples of images with negative sentiment and the second row reflects positive sentiment

The evaluation results are indicated in Table 2.6 in terms of accuracy, weighted accuracy, precision, and weighted precision metrics. The proposed $\mathcal{FSGDM}$ proves to be the outperforming model where it accomplishes an average accuracy of 75.53% and weighted accuracy of 77.27% against 63.56% and 63.55% for GDM using DAEM algorithm, compared to 62.15% and 62.13% for GDM using EM algorithm. This shows again the efficacy of the proposed model that takes into consideration the exact calculation of the Fisher matrix. Compared to DM using EM, DAEM, and even the Fisher scoring algorithm, it is clear how the generalized Dirichlet distribution tackles the problem better than other methods with reference also to Linear SVM and Logistic Regression [62].

Throughout all the conducted experiments, we prove the benefits of taking into account the full computation of the Fisher information matrix which demonstrated motivational effects on the

Table 2.6: Accuracy rates and precision results for SENTIBANK dataset

Models	Accuracy	W-Accuracy	Precision	W-Precision
Linear SVM [62]	67.00	-	-	-
Logistic Regression [62]	70.00	-	-	-
DM+EM	60.39	60.37	51.38	60.36
DM+DAEM	61.62	61.64	51.91	61.61
DM+FS	66.55	66.54	53.13	58.44
GDM+EM	62.15	62.13	51.15	62.14
GDM+DAEM	63.56	63.55	51.52	63.54
$\mathcal{FSGDM}$	75.53	77.27	51.32	75.46

detection of depression tweets, sentiment analysis, and emotion recognition. We compare also the computational cost of our proposed algorithm $\mathcal{FSGDM}$ with the GDM+DAEM in Table 2.7. The independence assumption in GDM gains time with regard to the calculation of the exact Fisher matrix as it is shown in Table 2.7. However, the calculation of the estimation of the parameters of $\mathcal{FSGDM}$ costs less in terms of memory as we expressed the derivative of the parameters using Beta-binomial probability density function.

Table 2.7: Comparing the performance of the proposed algorithm with the GDM approach in terms of memory usage and running time

	Memory usage (MB)		Running time (s)	
Dataset	GDM+DAEM	$\mathcal{FSGDM}$	GDM+DAEM	$\mathcal{FSGDM}$
Depression	15.75	12.14	16.24	12.34
Friends TV show	121.28	66.12	11.45	28.34
SENTIBANK	80.01	4.66	1.68	3.68

2.6 Conclusion

In many statistical models that concern count data, one common hypothesis adopted commonly is the independence assumption. Yet, in our work, we prove the benefits of taking into consideration the dependence and correlation of the feature vectors. For that purpose, we propose a new parametrization of the GDM and then we compute an exact Fisher information matrix. The exact calculation is based on the Beta-binomial probability distribution function and newly defined parameters. Such consideration was under the mixture of the GDM where we estimate the model's parameters through the Fisher scoring algorithm and optimally identified the number of clusters. With the proposed approach, we are able to reveal the importance of constructing count data modeling without the independence assumption. We implement the proposed $\mathcal{FSGDM}$ to three challenging applications: detecting depression in tweets, dialogue-based emotion recognition, and image-based sentiment analysis with three different modalities of count data namely text, dialogue, and images. For each of the mentioned applications, the proposed approach demonstrates robustness and high efficiency in terms of the obtained results. Future works could consider calculating the Fisher information matrix for other related count data models. In addition, a promising future work could concern a semi-supervised learning approach based on the developed model.

Chapter 3

Emotion recognition: A smoothed Dirichlet multinomial solution

Multinomial-based models have been extensively used for count data modeling and challenging applications such as image processing, text recognition, and behavioral sciences. Despite the good performance obtained with those models, they still suffer from challenging issues that require continuous exploring of other alternative approaches. In this work, we address the issue of smoothing language modeling. To the best of our knowledge, distributions defined in a smoothed simplex were not considered before as conjugate priors for the multinomial. We propose a smoothed Dirichlet multinomial (SDM) distribution and a mixture of SDMs with a likelihood-based learning. We evaluate the proposed approach on three challenging applications related to emotion recognition: depression on social media, happiness analysis, and pain estimation. The smoothed Dirichlet multinomial solution presents the best results comparing to the related works and the multinomial-based models such as Dirichlet compound multinomial and the multinomial model.

3.1 Introduction

Count data frequently occur in many domains such as machine learning, computer vision, psychology, behavioural sciences, and public health which make its analysis an essential task to detect abnormal behaviors, categorize text documents, and analyze emotion states. Considering the great challenges faced by count data modeling namely the sparseness, the high-dimensionality problem, and the overdispersion, numerous approaches have been proposed for this matter such as Hurdle models [5], zero-inflated [6], Negative-Binomial [7], and Poisson mixtures [63]. Basically, these techniques are not able to model counts effectively in real-world applications. For this fact, the most popular alternative count data modeling was the multinomial distribution.

Even though the accurate performance achieved with the multinomial distribution in several applications [9, 10, 64], this modeling suffers from multiple limitations such as the independence assumption [65, 54, 66]. This assumption simplifies the modeling of data but it is in fact erroneous in most real-world applications. Another problem with this model is the burstiness phenomenon [67] where the multinomial model fails to capture the words that have the probability of appearing once or the probability of seeing the event again. But, in reality if a word does appear once, it is much more likely to appear again.

Authors in [22] have proposed to represent a text document as a probability vector that leads to draw a scaled-bag-of-words from the Dirichlet distribution to become a bag-of-scaled document. The key difference with this model is that its parameters are not constrained to sum up to one which give the Dirichlet Compound Multinomial (DCM) extra degree of freedom. Thus, adding a prior probability for each word enables to combine information between the words by assuming that the probabilities of counts are related in a certain manner. When the parameters sum is larger than one, the counts become more bursty and less bursty when the sum is less than one. It has been shown in [22] that DCM is more appropriate for modeling text documents than the traditional multinomial model [68, 69]. In fact, the Dirichlet compound multinomial is capable of modeling the burstiness of words in text because the Dirichlet distribution has the form of $data^{parameter}$ that is able to take into account the power-law nature of text. However, the Dirichlet distribution has a restrictive negatively correlated covariance matrix. In addition, the Dirichlet is limited by reason of variables with the same mean must have the same covariance that makes the model insufficient and not suitable for challenging applications [70, 71].

Taking into account those disadvantages, the use of the generalized Dirichlet distribution [72] as an alternative was proposed in [27]. In fact, this distribution [73, 28, 74] allows to have more general covariance matrix to reflect different amounts of prior information which can be positively or negatively correlated. Further, the dependence among the posterior of the probabilities is slight and more flexible due to the extra parameters comparing to the Dirichlet. The generalized Dirichlet was not the only distribution proposed as a conjugate prior, but also the scaled Dirichlet was proposed which resulted in the multinomial scaled Dirichlet (MSD) [75]. In fact, the scaled Dirichlet [76, 77] is obtained from a perturbed random composition or a powering operation with a Dirichlet distribution. Further, Beta-Liouville distribution is a parametric generalization of the Dirichlet distribution and a natural choice for compositional data [78]. It belongs to the Liouville family of distributions where the generating density is chosen to be a Beta distribution. The Beta-Liouville is a conjugate prior to the multinomial distribution [79]. Based on this hypothesis, the author in [80] introduced the multinomial Beta-Liouville (MBL) distribution which presents an efficient count data modeling for accurate clustering applications.

Each distribution that has been presented previously as a conjugate prior is defined on the ordinary simplex $\Delta = \{\vec{X} = (x_1, \dots, x_D), x_d > 0, d = 1, \dots, D; \sum_{d=1}^{D} x_d = 1\}$. However, distributions that consider the whole simplex as its domain are not well defined since language models representing documents are generally smoothed and do not cover the whole domain. To overcome this issue, the count vector representing the text document is smoothly represented using Jelinek-Mercer smoothing technique [81]. In this regard, a smoothed Dirichlet distribution has been introduced in [82] which is a new form of the Dirichlet distribution defined on a smoothed simplex. For instance, it becomes interesting to introduce a new smoothed conjugate prior for the multinomial as a new count data modeling approach.

In this chapter, we propose a novel solution for emotion recognition where we adress count data challenging through the following contributions:

(1) Prove that Smoothed Dirichlet is a conjugate prior to the multinomial distribution.

(2) Propose a new count-data model namely smoothed Dirichlet multinomial mixtures with a likelihood-based learning.

(3) Detect emotional states by means of challenging applications such depression detection, psychology analysis, and pain estimation.

The rest of the chapter is organized as follows. We introduce in Section 2 the related multinomial priors that were previously proposed. In Section 3, we present the novel proposed multinomial-based model where we describe the smoothed Dirichlet multinomial mixture and its learning algorithm. Section 4 displays the considered emotion recognition experiments including depression on social media, psychology analysis, and pain estimation. Section 5 concludes the paper with potential future works.

3.2 Related work

Since our work is related to proposing a new multinomial prior for emotion recognition, we introduce the previously proposed priors, present the mathematical background of each prior distribution, and review the recent related-works proposed for emotion recognition.

3.2.1 Multinomial priors

Multinomial distribution has been extensively used for count data modeling and for text recognition especially. Though the fame and the well-known properties of this model, it suffers from a variety of issues such as the independence assumption, the burstiness problem and the ineffectiveness to handle the sparsity nature of data. Taking these challenges into consideration, smoothing the parameter of the multinomial distribution using a conjugate prior was an effective solution. As first thought, the Dirichlet distribution which is known as the most convenient conjugate prior for the multinomial was proposed as a smoothing distribution that gives the Dirichlet Compound Multinomial (DCM) [22].

$$\begin{aligned} p(\vec{X}|\vec{\alpha}) &= \int_\theta p(\vec{X}|\vec{\theta})p(\vec{\theta}|\vec{\alpha})d\theta \\ &= \frac{\Gamma(\sum_{d=1}^{D} x_d + 1)\Gamma(\sum_{d=1}^{D} \alpha_d)}{\Gamma(\sum_{d=1}^{D} x_d + \alpha_d)} \prod_{d=1}^{D} \frac{\Gamma(x_d + \alpha_d)}{\Gamma(\alpha_d)\Gamma(x_d + 1)} \end{aligned} \tag{47}$$

where $\vec{X}$ is a count vector, $p(\vec{X}|\vec{\theta})$ is a multinomial distribution that depends on D parameters $\theta_1, \ldots, \theta_D$, $p(\vec{\theta}|\vec{\alpha})$ is the conjugate Dirichlet prior defined with an $\vec{\alpha}$ parameter.

Actually, although the DCM has achieved advantageous performance in various applications [11, 25, 23], the Dirichlet prior suffers from some limitations such as the restrictive covariance between two random variables that have to be negatively correlated and the proportional relation between the variance and the mean for two different variables. Proposing a general distribution that is able to handle these disadvantages was the purpose of the generalized Dirichlet distribution. Indeed, this generalized distribution is also a prior for the multinomial which makes the introduction of generalized Dirichlet multinomial (GDM) distribution [27].

$$\begin{aligned} p(\vec{X}|\vec{\alpha},\vec{\beta}) &= \int_{\theta} p(\vec{X}|\vec{\theta})p(\vec{\theta}|\vec{\alpha},\vec{\beta})d\theta \\ &= \frac{\Gamma(\sum_{d=1}^{D} x_d + 1)}{\prod_{d=1}^{D}\Gamma(x_d+1)} \prod_{d=1}^{D-1} \frac{\Gamma(\alpha_d+\beta_d)}{\Gamma(\alpha_d)\Gamma(\beta_d)} \prod_{d=1}^{D-1} \frac{\Gamma(\alpha'_d)\Gamma(\beta'_d)}{\Gamma(\alpha'_d+\beta'_d)} \end{aligned} \tag{48}$$

where $p(\vec{\theta}|\vec{\alpha},\vec{\beta})$ is the conjugate generalized Dirichlet prior with parameter vector $(\alpha_1, \beta_1, \ldots, \alpha_{D-1}, \beta_{D-1})$, $\alpha'_d = \alpha_d + x_d$, and $\beta'_d = \beta_d + x_{d+1} + \cdots + x_D$, for $d = 1, \ldots, D-1$.

Even though the high performance achieved by MGD mixtures involving different applications (spatial colour image indexing, handwritten digit recognition, and text document clustering), it requires learning large numbers of parameters that increase the model's complexity and the running time. One more alternative to the Dirichlet distribution as a conjugate to the multinomial is the scaled Dirichlet (SD). The SD is a generalization of the Dirichlet by perturbation-combination operation and has more parameters (extra D degrees of freedom) which makes the distribution more flexible and removes the requirement of equally scaled parameter in Dirichlet distribution. In case of using this distribution as a prior to the multinomial, the multinomial scaled Dirichlet mixture model was introduced to recognize text documents in [75].

$$
\begin{aligned}
p(\vec{X}|\vec{\alpha},\vec{\beta}) &= \int_{\theta} p(\vec{X}|\vec{\theta})p(\vec{\theta}|\vec{\alpha},\vec{\beta})d\theta \\
&= \frac{\Gamma(\sum_{d=1}^{D} x_d + 1)}{\prod_{d=1}^{D} x_d} \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)}{\Gamma(\sum_{d=1}^{D} x_d + \alpha_d) \prod_{d=1}^{D} \beta_d^{x_d}} \prod_{d=1}^{D} \frac{\Gamma(\alpha_d + x_d)}{\Gamma(\alpha_d)}
\end{aligned} \tag{49}
$$

where $p(\vec{\theta}|\vec{\alpha},\vec{\beta})$ is the conjugate scaled Dirichlet prior with parameter vector $(\alpha_1, \beta_1, \dots, \alpha_D, \beta_D)$.

Although the good modeling flexibility obtained when using this model, there is a need to learn large number of parameters as MGD. Moreover, the scaled Dirichlet has no closed form for the mean, the variance, and the covariance which restrict the use of this distribution in such complicated applications. All these drawbacks can be handled by using the Beta-Liouville (BL) distribution as an alternative prior for the multinomial distribution. The BL distribution has more general covariance structure. The composition of the BL and the multinomial gives the multinomial Beta-Liouville distribution in [80].

$$
\begin{aligned}
p(\vec{X}|\vec{\alpha},\vec{\beta}) &= \int_{\theta} p(\vec{X}|\vec{\theta})p(\vec{\theta}|\vec{\alpha},\vec{\beta})d\theta \\
&= \frac{\Gamma(\sum_{d=1}^{D} x_d + 1)}{\prod_{d=1}^{D} \Gamma(x_d + 1)} \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)\Gamma(\alpha')\Gamma(\beta') \prod_{d=1}^{D-1} \Gamma(\alpha_d')}{\Gamma(\sum_{d=1}^{D} \alpha_d')\Gamma(\alpha'+\beta')\Gamma(\alpha)\Gamma(\beta) \prod_{d=1}^{D-1} \Gamma(\alpha_d)}
\end{aligned} \tag{50}
$$

where $p(\vec{\theta}|\vec{\alpha},\vec{\beta})$ is the conjugate Beta-Liouville prior with parameter vector $(\alpha_1, \beta_1, \dots, \alpha_{D-1}, \alpha, \beta)$, $\alpha_d' = \alpha_d + x_d$, $\alpha' = \alpha + \sum_{d=1}^{D-1} x_d$, and $\beta' = \beta + x_D$, for $d = 1, \dots, D-1$.

3.2.2 Emotion recognition

Distress, sadness, or any other emotions could be inflicted by erroneous habits such as spending more time connecting electronically than spending time with family and friends or due to unexpected circumstances as earthquakes, floods, wars, or else by virtue of infectious diseases like the COVID-19. Human emotional states are altered with delightful and dejected moments which can be expressed through the facial muscles or influence the content of social media status (text, image, video). The later could be characterized by short texts which suffer from the problem of sparseness or facial expressions that exhibit short and long spatio-temporal variations. In this context, different machine learning methods and deep learning algorithms have been proposed to address these challenges. Earlier works were concentrated on single modality such as facial expressions extracted from images and videos where the recent focus was shifted now to multi-modal and context-aware emotion recognition as proposed by [83, 84, 85, 86, 87, 88]. [84] proposed a three-stream CNN models. The first stream is for body feature extraction, the second encodes context features, and the last one performs as a fusion network combining the features of the two CNNs. A similar network architecture was introduced by [86], the CAER-Net, a two stream encoding network which includes facial-stream and context-stream. A Region Proposal Network (RPN) was proposed by [85] to detect context elements integrated into a branch of Graph Convolutional Network (GCN). Another parallel branch of CNN was incorporated for extracting body features. The work of [88] presents a multi-modal and context-aware emotion recognition framework that combines three interpretations of context from psychology principles. The first context uses multiple modalities of faces and gaits, the second encodes semantic information using self-attention-based CNN and the last one concerns socio-dynamic inter-agent interactions through depth maps. [87] demonstrates the utility of multi-task CNN for multi-modal context-based emotion recognition. Their proposed framework includes three modules: body features extraction, scene features extraction, fusion and decision module.

3.3 Smoothed Dirichlet Multinomial

The smoothed Dirichlet (SD) [82] distribution was proposed as an approximation to the Dirichlet distribution defined in a smoothed simplex. It was empirically demonstrated that SD distribution outperforms multinomial and DCM models in text classification task. Generating smoothed document representation is the core idea of the SD distribution where a smoothed proportion x^s is defined as follows:

$$x^s = \lambda x^u + (1-\lambda)x^{ge} \tag{51}$$

where $0 < \lambda < 1$ is a smoothing parameter, x^u is a proportional vector, and x^{ge} is the general proportion of all the words presented in a text document [89]. Based on the generation of this smoothing proportions, a smoothed subset of a unit simplex $\Delta = \{\vec{X} = (x_1, \ldots, x_D), x_d > 0, d = 1, \ldots, D; \sum_{d=1}^{D} x_d = 1\}$ is defined as follows:

$$\Delta^s = \{\vec{X}^s\} = \{\lambda\vec{X}^u + (1-\lambda)\vec{X}^{ge} | \vec{X}^u \in \Delta\} \tag{52}$$

The smoothed Dirichlet distribution has the same parametric form as the Dirichlet distribution but defined on the smoothed simplex Δ^s as follows:

$$p(\vec{X}^s|\vec{\alpha}) = \frac{S^S}{\prod_{d=1}^{D} \alpha_d^{\alpha_d}} \prod_{d=1}^{D} (x_d^s)^{\alpha_d - 1} \tag{53}$$

where $S = \sum_{d=1}^{D} \alpha_d$ and $\alpha_d > 0,\ d = 1, \ldots, D$ is the shape parameter that characterizes the SD distribution.

In this section, we prove that SD distribution is a conjugate prior for multinomial distribution. Consider the joint distribution of $\vec{X}$ and $\vec{\theta}$ and the probability $p(\vec{X}|\vec{\theta})$ of the count vector $\vec{X}$ given the parameter $\vec{\theta}$ is a multinomial distribution and the prior $p(\vec{\theta}|\vec{\alpha})$ is a smoothed Dirichlet given as follows:

$$\begin{aligned} p(\vec{X}, \vec{\theta}|\vec{\alpha}) &= p(\vec{X}|\vec{\theta})p(\vec{\theta}|\vec{\alpha}) \qquad (54) \\ &= \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D} (\theta_d^s)^{x_d} \frac{S^S}{\prod_{d=1}^{D} \alpha_d^{\alpha_d}} \prod_{d=1}^{D} (\theta_d^s)^{\alpha_d - 1} \\ &= \frac{|x|!}{\prod_{d=1}^{D} x_d!} \frac{S^S}{\prod_{d=1}^{D} \alpha_d^{\alpha_d}} \prod_{d=1}^{D} (\theta_d^s)^{x_d + \alpha_d - 1} \end{aligned}$$
$$\tag{55}$$

Then, the posterior is given by:

$$
\begin{aligned}
p(\vec{\theta}|\vec{X},\vec{\alpha}) &= \frac{p(\vec{X},\vec{\theta}|\vec{\alpha})}{p(\vec{X}|\vec{\alpha})} \\
&= \frac{\sum_{d=1}^{D}(x_d+\alpha_d)^{S+|x|}}{\prod_{d=1}^{D}(x_d+\alpha_d)^{x_d+\alpha_d}}\prod_{d=1}^{D}(\theta_d^s)^{x_d+\alpha_d-1}
\end{aligned}
\qquad (56)
$$

which is a smoothed Dirichlet distribution with parameters $(x_1+\alpha_1,\ldots,x_D+\alpha_D)$. Integrating the joint distribution over $\vec{\theta}$, we obtain the following new distribution so-called smoothed Dirichlet multinomial (SDM) distribution:

$$
\begin{aligned}
p(\vec{X}|\vec{\alpha}) &= \int_\theta p(\vec{X},\vec{\theta}|\vec{\alpha})d\theta \\
&= \frac{|x|!}{\prod_{d=1}^{D}x_d!}\frac{S^S}{\prod_{d=1}^{D}\alpha_d^{\alpha_d}}\int_\theta\prod_{d=1}^{D}(\theta_d^s)^{x_d+\alpha_d-1}d\theta \\
&= \frac{|x|!}{\prod_{d=1}^{D}x_d!}\frac{S^S}{\prod_{d=1}^{D}\alpha_d^{\alpha_d}}\frac{\prod_{d=1}^{D}(x_d+\alpha_d)^{x_d+\alpha_d}}{\sum_{d=1}^{D}(x_d+\alpha_d)^{S+|x|}}
\end{aligned}
\qquad (57)
$$

where $|x|=\sum_{d=1}^{D}x_d$.

The availability of challenging applications increasingly faced with multimodal data gives rise to the deployment of mixture models. Given their powerful properties, we propose a mixture of SDM distributions given by:

$$
p(\vec{X}|\Theta)=\sum_{j=1}^{K}p_j p(\vec{X}|\vec{\alpha}_j) \qquad (58)
$$

where p_j is the mixing weight, $p(\vec{X}|\vec{\alpha}_j)$ is the j-th component refers to SDM, and $\Theta=(\vec{\alpha}_1,\ldots,\vec{\alpha}_K, p_1,\ldots,p_K)$ is the entire set of parameters.

Given a set of N count data vectors $\mathcal{X}=(\vec{X}_1,\ldots,\vec{X}_N)$, the log-likelihood of SDM mixtures is defined as:

$$\mathcal{L}(\Theta, \mathcal{X}) = \log \prod_{i=1}^{N} p(\vec{X}_i|\Theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} p_j p(\vec{X}_i|\vec{\alpha}_j) \tag{59}$$

For estimating SDM mixture model's parameters, we consider a maximum-likelihood approach on the basis of the Expectation-Maximization algorithm (EM) [47]. However, the estimation of Θ parameters does not present a closed-form solution. Thus, we employ for that a Newton-Raphson method given by:

for $d = 1, \ldots, D, j = 1, \ldots, K$

$$\hat{\alpha}_{jd} = \alpha_{jd} - \frac{f(\alpha_{jd})}{f'(\alpha_{jd})} \tag{60}$$

where

$$\begin{aligned} f(\alpha_{jd}) &= \frac{\partial \mathcal{L}(\Theta, \mathcal{X})}{\partial \alpha_{jd}} \qquad (61) \\ &= \frac{\partial}{\partial \alpha_{jd}} \sum_{i=1}^{N} r(i,j) \Big[\log(|x|!) - \sum_{d=1}^{D} \log(x_d!) \\ &+ S \log S - \sum_{d=1}^{D} \alpha_{jd} \log \alpha_{jd} + \sum_{d=1}^{D} (\alpha_{jd} + x_{id}) \log(\alpha_{jd} + x_{id}) \\ &- (S + |x|) \log(S + |x|) \Big] \\ &= \sum_{i=1}^{N} r(i,j) \Big[\log \frac{S}{\alpha_{jd}} + \log \frac{x_{jd} + \alpha_{jd}}{S + |x|} \Big] \end{aligned}$$

where $r(i,j) = p_j p(\vec{X}_i|\vec{\alpha}_j) / \sum_{j=1}^{K} p_j p(\vec{X}_i|\vec{\alpha}_j)$, and:

$$f'(\alpha_{jd}) = \sum_{i=1}^{N} r(i,j) \Big[\frac{1}{S} - \frac{1}{\alpha_{jd}} + \frac{1}{x_{id} + \alpha_{jd}} - \frac{1}{S + |x|} \Big] \tag{62}$$

The complete learning algorithm is given in algorithm 2.

Algorithm 2: Smoothed Dirichlet multinomial learning algorithm

```
Input: $\mathcal{X}$ dataset, Number of components $K$ ;
Output: $\Theta^*$ ;
Initialization using K-means and the Method of Moment as in [90];
repeat
    foreach Component $j$ do
        Estimate the posterior distribution $r(i, j)$ using Bayes rule ;
        Estimate the mixing weight components $p_j = \frac{1}{N}\sum_{i=1}^{N} r(i, j)$ ;
        Update the parameters $\alpha_{jd}$ $(d = 1, \ldots, D)$ using Newton-Raphson method ;
    end
until Convergence of Likelihood;
```

3.4 Experiments analysis: emotion recognition

In this section, we validate our proposed Smoothed Dirichlet multinomial on different emotion recognition applications. We consider detecting depression on social media, analyzing happiness intensity, and estimating pain levels.

3.4.1 Depression on social media

Social media have been always considered as a double-edged sword and could be viewed as a first reason that affect teenager emotional states. Currently, teenager and young adult spend the majority of their daily time connecting electronically through social media which affects their relationship to their peers, family, and friends. This fact leaves them socially isolated with poorer concentration, self deprivation, and depression [91, 92, 93]. A clinical psychologist at the Child Mind Institute states “The less you are connected with human beings in a deep, empathic way, the less you’re really getting the benefits of a social interaction,” [1]. In this context, we are interested to detect early depression states on social media through a tweet dataset [2] which is constructed from two separated datasets: the well-known Sentiment140 dataset which is highly imbalanced [94] and depressive tweets generated from Twint tool. The depression dataset contains 10,314 tweets distributed into positive tweets extracted from Sentiment140 and depressive tweets produced from Twint [3]. We generate a bag-of-words from the totality of the tweets and we remove the stop, rare,

[1] https://childmind.org/article/is-social-media-use-causing-depression/

[2] https://github.com/viritaromero/Detecting-Depression-in-Tweets

[3] https://github.com/twintproject/twint

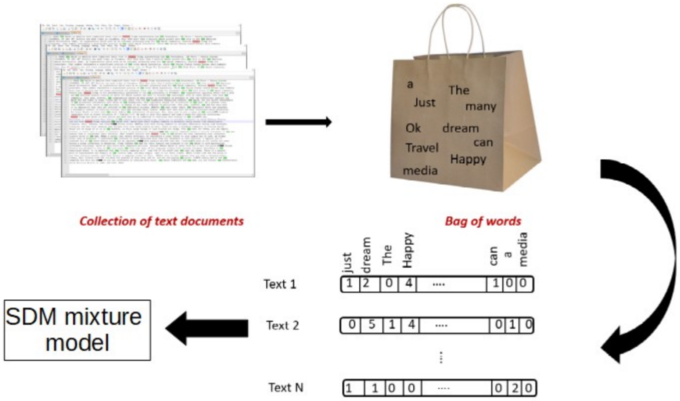


Figure 3.1: Illustration of depression recognition approach

and short words with less than 10 occurrences. We explain in details the generation of bag-of-words from depression tweets and the adaption to SDM model in Figure B.1.

Table 3.1: Comparison between different multinomial-based models on Depression-Tweets dataset

Model	Accuracy	Precision	Recall	F-measure
Multinomial	62.35	64.44	43.56	51.93
DCM	65.37	50.89	50.87	50.88
DCM+DAEM	66.27	52.47	52.38	52.42
SD	66.11	48.87	40.24	42.87
SDM	69.68	71.01	46.73	56.37
SDM + noise	68.78	83.13	40.12	54.13

We evaluate the obtained results using different metrics: accuracy, precision, recall, and F-measure [95] and we compare the experimental results of the SDM with different multinomial-based models on the same Depression-Tweets dataset. We clearly observe from Table 3.1 that SDM mixture model outperforms the multinomial, DCM, and the smoothed Dirichlet (SD). The SDM achieves 69.68% when detecting the depression tweets with a 71.01% of precision that presents more that 10% of difference between the other related multinomial-based models. We add also a Gaussian noise to the data to quantify the robustness of SDM algorithm. We note that we obtain almost the same performance which proves more the efficiency of the proposed model. We mention that the Dirichlet compound multinomial has been proposed also with a different learning method, the deterministic-annealing EM (DAEM) [96] that makes an improvement with regards to the DCM with EM. However, even we present our SDM solution with an EM learning method, it is shown that it outperforms all the related models which opens a new direction for an improved SDM with a DAEM learning model as a future direction.

3.4.2 Psychology analysis

As we mentioned before, emotions could be expressed through texting on social media or using facial expressions detected by means of images and videos. Humans are able to express more than 10,000 expressions through 43 facial muscles which makes reading faces a significant human skill and a challenge task for artificial intelligence algorithms. Researchers generally decode the human face into six basic expressions: disgust, happiness, fear, anger, sad, and surprise. For this purpose, we consider the EMOTIC dataset [97, 84] which contains 26 emotional expressions with a total of 23,571 images. In this work, we are interested to recognize different expressions that are associated with positive emotions. Thus, we extract a subset EMOTIC-PA from the EMOTIC dataset that are related to analyze positive emotions intensity including: affection, excitement, happiness, peace, pleasure, and sympathy as shown in Figure 3.2. Emotions in the EMOTIC dataset present several recognition challenges as the images are annotated by five different annotates and represented using the continuous dimensions of the VAD Emotional State (V: Valence which encodes the happiness and the pleasure and ranges from positively to negatively, A: Arousal that measures the level of human agitation ranging from excitement to relaxation, D: Dominance which estimates the control level and ranges from submission to dominant). For inter-rater reliability purposes, authors [84] computed an agreement score between all the annotators using the average of Fleiss' Kappa values for each category and the average standard deviation for the continuous dimensions. For the sake of representing the images as count data vectors, we extract first SIFT features from 16×16 patches on regular grids of 8 pixels. Then, we construct a bag-of-visual words using the K-means clustering algorithm where we generate different size of visual vocabulary ranging from 100 to 2500. We illustrate the steps of psychology analysis approach in Figure 3.3.

We investigate the performance of the smoothed Dirichlet multinomial using two metrics: the mean average precision (mAP) and the accuracy. We compare the obtained results with the related multinomial-based models and the algorithms that have been applied on the EMOTIC dataset for

Figure 3.2: Visual examples of the six feeling from the EMOTIC database for the Psychology analysis

Table 3.2: Psychology analysis using different algorithms (mAP)

	Affection	Excitement	Happiness	Peace	Pleasure	Sympathy
[84]	27.85	77.16	58.26	21.56	45.46	14.71
[85]	46.89	71.89	73.26	32.85	57.46	17.53
[86]	19.9	35.26	49.26	16.72	19.47	17.12
[88]	36.78	82.21	68.21	35.14	61.34	24.63
[87]	13.42	47.94	47.95	14.62	28.98	8.46
Multinomial	33.92	35.14	35.16	31.56	32.55	31.48
DCM	50.19	83.33	47.73	36.11	41.67	46.04
SD	50.88	52.89	53.38	49.12	50.01	61.11
SDM	60.01	56.88	56.35	57.89	59.30	50.01

the selected emotions (affection, excitement, happiness, peace, pleasure, and sympathy). We mention in Table 3.2 that our proposed SDM outperforms the other related-models and presents equally distributed mean average precision for all the emotions intensities contrary to the other models proposed in literature such as [84, 85, 86, 87, 88] that create an imbalance in recognized emotions.

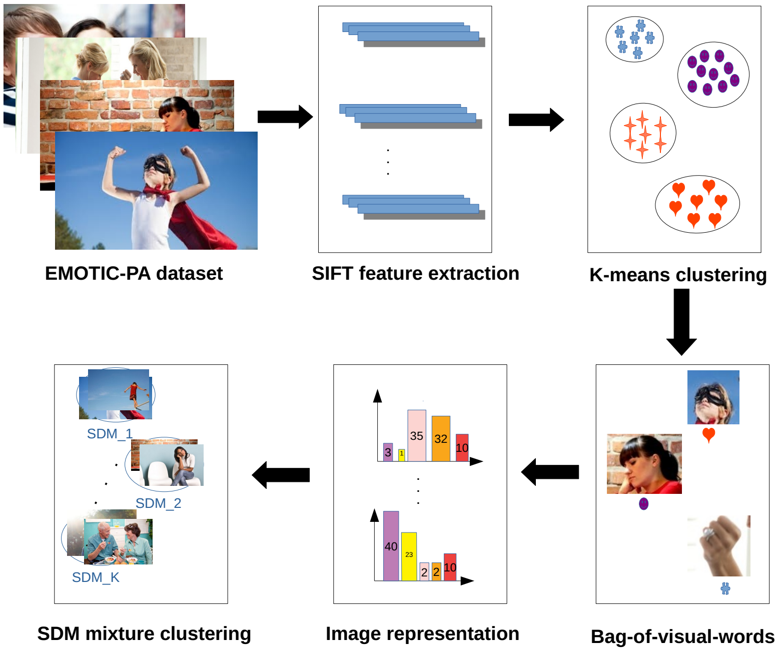


Figure 3.3: Description of psychology analysis approach using SDM mixture-based clustering

Table 3.3 clearly shows the outperformance of the SDM model as compared to the multinomial-based models in terms of accuracy results when comparing the mean average precision. Our proposed approach is based on multinomial distribution which is able to adapt count data presented by bag-of-visual words contrary to the compared related works that are based on neural networking method.

We compare the performance of the proposed approach with the related multinomial-based models in Figure 3.4 where we evaluate the memory occupied and the time of execution with regards to the features dimensions. We mention that the fastest algorithm is the smoothed Dirichlet and the multinomial distribution while the DCM is the slowest approach. In terms of memory, the SDM model occupies less than the other models which proves again the efficiency of the proposed algorithm.

3.4.3 Pain detection

Facial pain detection [98, 99] has been recently an emergent research area related to several health care applications. Indeed, determining the pain level is a challenging task even for clinical professors. The Action Units which have been generally used for facial expression detection

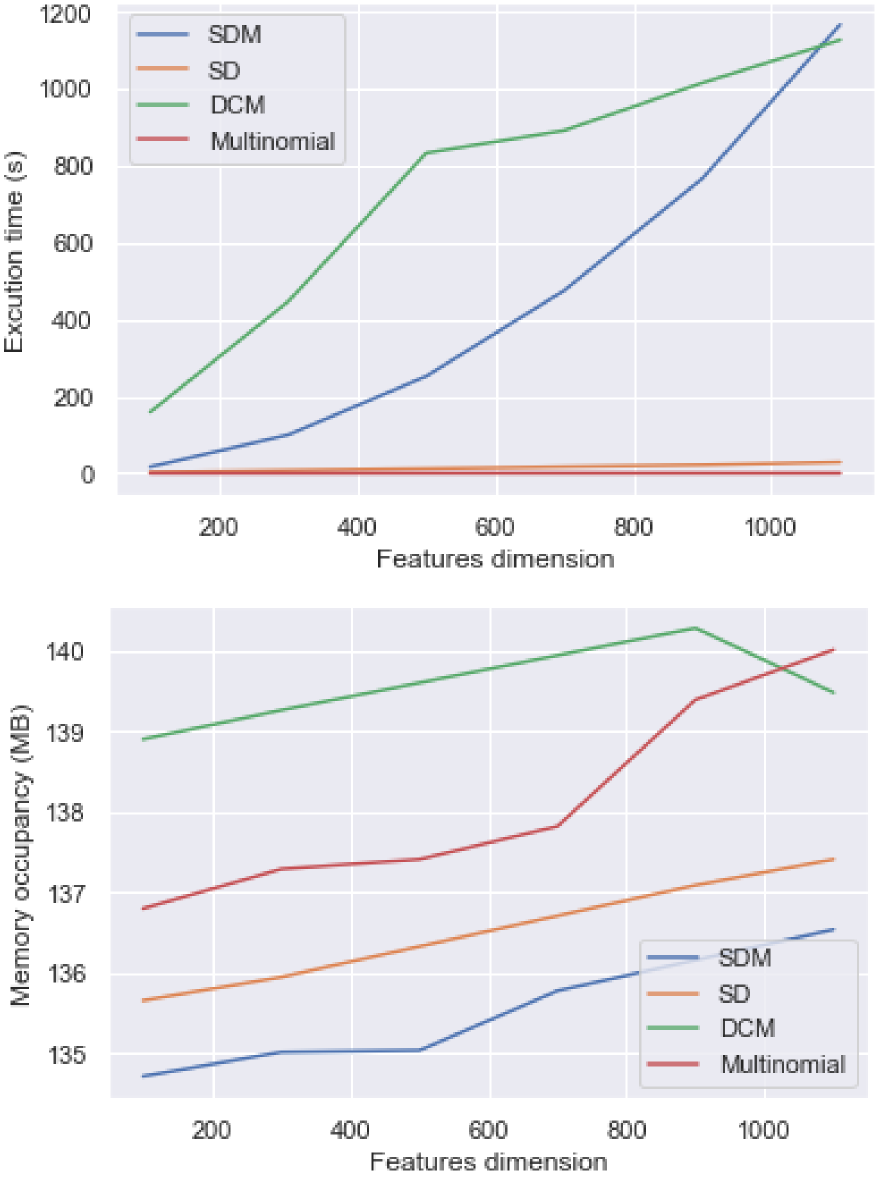


Figure 3.4: Performance evaluation in terms of memory usage and running time for psychology analysis

Table 3.3: Evaluation results on EMOTIC-PA dataset

Models	Accuracy	mAP (%)
[84]	-	40.83
[85]	-	49.98
[86]	-	26.28
[87]	-	26.89
[88]	-	51.38
Multinomial	32.65	33.92
DCM	34.36	50.84
SD	34.43	53.47
SDM	34.17	57.31

are not sufficient to detect the different levels of pain from images. For this purpose, we consider the BioVid Heat pain dataset [100] which selects emotion features related to pain recognition using multiple "Kinect" camera, highly-controlled pain stimulation, and simultaneous data collected from skin conductance level (SCL), electrocardiogram (ECG), electromyogram (EMG), and electroencephalography (EEG). The dataset (Figure 3.5) consists of a total of 90 participants (men and women) from three ranges of age groups: $18-35$, $36-50$, and $51-65$. In this paper, we consider only part A of the dataset which is a subsection containing only video sequences of facial pain levels with a total of $8,700$ sequences where there are 100 samples per person making 1,740 videos per intensity levels. We extract LBP features from each frame and we construct from the extracted features a bag-of-visual words.

The smoothed distribution presents the best results comparing to the related works and the multinomial-based models such as DCM and the multinomial model. It is clearly noticeable in Table 3.4 that SD and SDM achieve outstanding results with an increasing of 10% in the accuracy. In this challenging application, the smoothed Dirichlet multinomial presents comparable results with the SD where we present also the results obtained in the state-of-the-art using different descriptors LBP, LPQ, and BSIF with SVM classifier [101]. We mention that our proposed approach employs LBP for feature extraction which outperforms all the related methods using as a classifier SVM applied for the BioVid-Heat Pain dataset.

We compare also the computational cost of our proposed algorithm with the multinomial-based

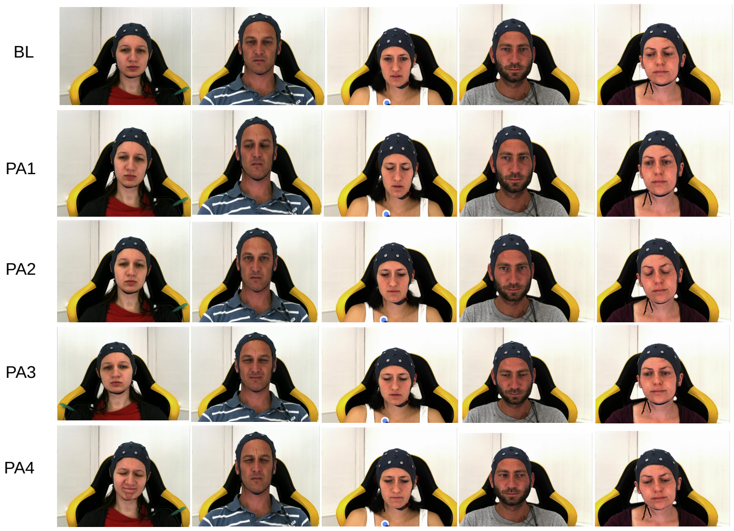


Figure 3.5: Face samples from BioVid-Heat Pain dataset accross pain intensities (PA*) and baseline (BL)

models in Figure 3.6. We evaluate the scalability of SDM with regards to dataset size. We report the results in terms of different number of frames of BioVid video sequences where we have: 20 frames (880 images), 40 frames (3280 images), 60 frames (4845 images), 80 frames (6445 images), 100 frames (8045 images), and 138 frames (11005 images). The parameters learning in DCM gains more time compared to SD, SDM, and multinomial inference. Further, the execution of SDM algorithm for pain detection costs the least in terms of memory.

Table 3.4: Accuracy results for Pain detection on BioVid-Heat Pain dataset

Method	Accuracy
LBP	59.08
LPQ	58.82
BSIF	59.25
LBP + LPQ	60.23
LPQ + BSIF	59.83
[101]	60.23
Multinomial	55.88
DCM	56.14
SD	69.60
SDM	67.11

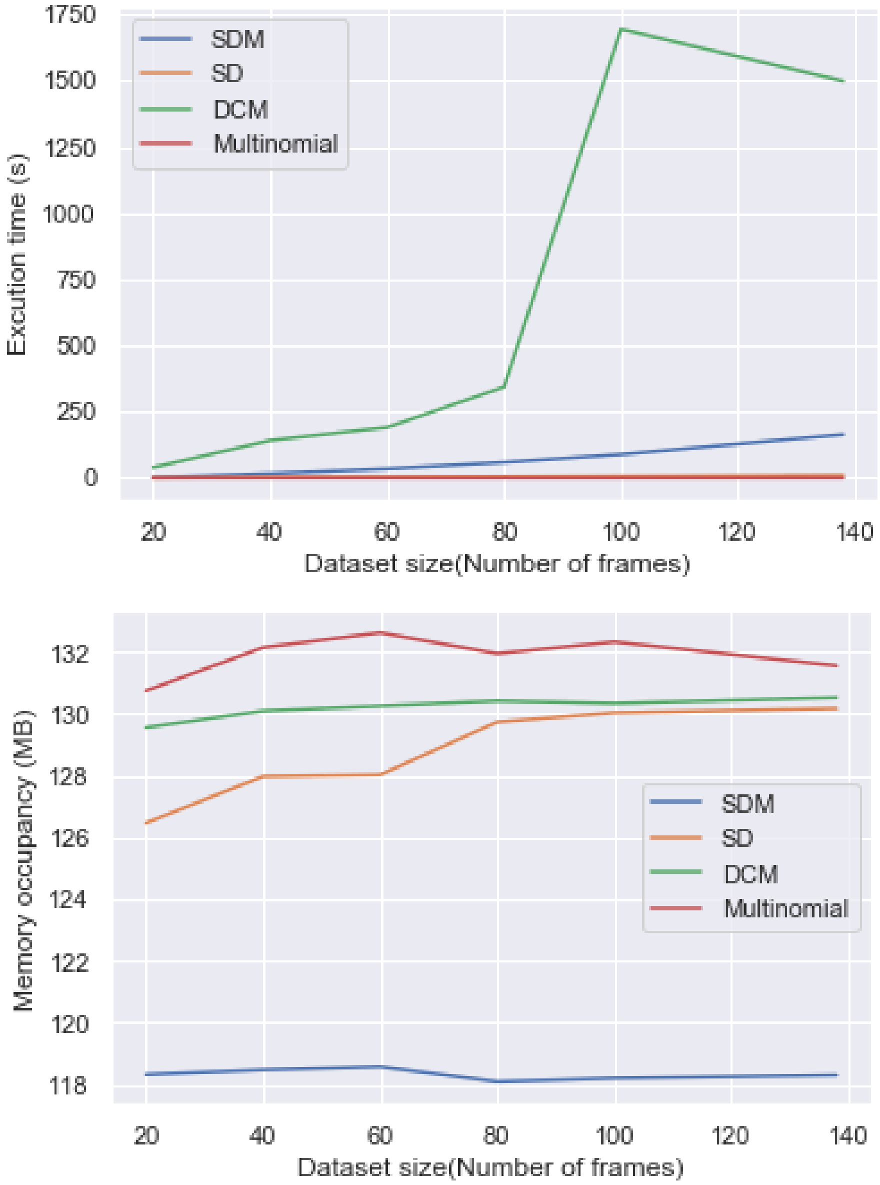


Figure 3.6: Performance evaluation in terms of memory usage and running time for Pain detection

3.5 Conclusion

We proposed, in this work, a novel multinomial-based model namely the smoothed Dirichlet multinomial distribution. The proposed SDM is defined on a smoothed simplex where the conjugate prior for the multinomial defines smoothed count vectors. We introduced a mixture of SDMs and the learning algorithm for the mixture parameters. The proposed approach was considered as an emotion recognition solution where we addressed the problem of three challenging applications: depression on social media, psychology analysis, and estimating pain levels. The obtained results show the outperformance of the novel SDM model with regards to the other related multinomial-based models. The obtained promising results open other directions for considering new conjugate priors for the multinomial distribution defined on a smoothed simplex such as Generalized Dirichlet and Beta-Liouville. Besides, it is interesting to address other challenging topics with the smoothed Dirichlet multinomial solution. For instance, it is worth noting to consider SDM approach for anomaly detection in video surveillance and medical image processing as the approach are able to adapt very well the same type of data.

Chapter 4

Smoothed Generalized Dirichlet: a novel count data model for detecting emotional states

In this chapter, we propose novel approaches to deal with the problem of burstiness, the challenge of count data sparseness, and the curse of dimensionality. We introduce a Smoothed Generalized Dirichlet distribution that is a smoothed variant of the generalized Dirichlet distribution and a generalization of the smoothed Dirichlet. We provide different learning methods based on mixture models and agglomerative clustering-based geometrical information: Kulback-Leibler, Fisher metric, and Bhattacharyya distance. Moreover, we show that the new Smoothed Generalized Dirichlet could be considered as a prior to the multinomial which generates a new distribution for count data that we call the Smoothed Generalized Dirichlet multinomial. In particular, we present an approximation based on Taylor series expansion for better performance and optimized running time in the case of high-dimensional count data. Proposed models are evaluated through two emotion detection applications: disaster tweets related emotions and pain intensity estimation. Experiments show the efficiency and the robustness of our approaches when dealing with texts, videos, and images.

4.1 Introduction

Counts, typically emanate from a variety of observations we make through the world around us. Observations could be trivial events in our daily life, for example, employee absences, number of traffic accidents per day, or even number of children ever born. Understanding this kind of observations is a high critical issue in analyzing infectious diseases, interpreting insurance data, categorizing topics in text documents, forecasting economic risks, and retrieving information from multimedia data. The analysis of count data considering the concurrency nature of words in a collection of text documents or visual words when considering multimedia datasets leads to challenges related to sparseness, curse of dimensionality, and burstiness. Sparseness, or the phenomenon of data with excess of zeros, appears whenever a word has not occurred in a document when considering a "bag-of-words" structure. The burstiness phenomenon [3] described as an accidental repetition of infrequent words or phrases in long documents was closely related to the "document-level burstiness" and "within-document burstiness" which is somehow a consequence of the overused conditional assumption of independence. In a different viewing, the burstiness was introduced in the work of Church & Gale [4] with the assumption of "bag-of-words" where the bursty occurrences were depicted using the frequencies of words. In a such meaning, the burstiness is postrayed as a contagious disease where if a few instances of a rare word have already occurred in a document, then there is a great probability to have some more instances of it. These phenomena are subjects of great interest for many research works as in [22, 102, 82, 11, 103, 104, 105, 106, 66, 107].

Choosing the adequate distribution that fits count data presents a considerable challenge for researchers as it is positively skewed with a high frequency of zeros values. Considering this fact, Hurdle [5] and zero-inflated models [6, 108] have been proposed for count data to handle the sparsity problem. The basic principle of hurdle models is based on Bernoulli distributions while for the zero-inflated, the count variables are modeled by a mixture of distributions supporting the non-negative integers such as Poisson and Negative Binomial [109]. Bernoulli models [110] have been utilized for analyzing binary behavior, for example, smoker-non smoker, vegan-non vegan, fail-succeed, and other outcomes that need to be 0 or 1. For all that, the exigency to figure out the number of times an event occurs makes the Poisson distribution [4] popular to incorporate term frequency information.

Indeed, a crucial problem that occurs when modeling count data is overdispersion which means that the variance of data is larger than the expected value. Overdispersion occurs due to the heterogeneity (non uniform) of data which is contrary to assumptions within simple parametric models. To deal with such defect, the Negative Binomial [7] has been considered instead. Rather than modeling the frequencies of words with one parameter by Poisson distribution, the Negative Binomial has two parameters specifying the number of failures and the success probability. Considering challenging applications such as language modeling, pattern recognition, and computer vision, these simple parametric models are unable to accurately fit count data.

The primary motivation of this work is to propose novel approaches for count data modeling that address simultaneously the problems of sparseness, burstiness, overdispersion, and high-dimensionality.

The remainder of this chapter is organized as follows. Section 2 presents the existing literature related to count data modeling and emotion recognition. In Section 3, we give a background introduction for smoothed Dirichlet and generalized Dirichlet distributions. In Section 4, we describe the text generator distribution that is proposed to deal with the sparseness and overdispersion issues. We present in Section 5 the parameters estimation algorithm, and the clustering mechanisms using mixture models and geometrical information. After, we provide a new multinomial prior in Section 6 addressing the challenge of burstiness and the overdispersion phenomena. Then, we present the obtained results of our proposed approaches through two applications namely recognizing disaster tweets related emotions and pain intensity estimation in Sections 7 and 8. Further, in subsection 8.3, we give an approximation to deal with the problem of high-dimensional count data. We finally conclude the paper in Section 9 along with open research questions.

4.2 Related work

This chapter builds on a long line of works regarding count data modeling and emotion recognition. As far as our contributions concern proposing new count data models and address the problem of emotion categorization, we review separately the previous works on both topics.

4.2.1 Emotion categorization algorithms

Understanding human emotions is a key factor for developing intelligent machines [111, 112]. Even though much research works have been proposed for the field of sentiment analysis and emotion recognition, it continues to present considerable challenges and new directions such as understanding cause of sentiment, sentiment dialogue generation and sentiment reasoning [113]. Emotion recognition is typically based on human face landmarks, speech signals, and text messages. In this work, we are interested in the literature of emotion categorization from texts, images and videos data. Recent approaches for sentiment analysis that employ machine learning have been mostly based on deep learning algorithms and contextual language models such as BERT [114], SenticNet [115], and RoBERTa [116]. Early deep learning studies on sentiment analysis have considered attention mechanism as Attention-based Bidirectional CNN-RNN proposed in [117], Contextual Attention-based LSTM introduced to model contextual relationships [118] and Multi-task Multimodal Emotion and Sentiment (MMES) that considers contextual inter-modal attention Framework [119]. To improve the performance of emotion recognition models, a multitask-based framework was combined with deep learning network [120]. A top-down and bottom-up learning technique using an ensemble of symbolic and subsymbolic AI tools was proposed in [121] for sentiment analysis. Other than deep learning and contextual language models, reinforcement learning approaches were also considered for emotion and sentiment analysis such as the work of [122] which considers multimodal context within domain knowledge. Regarding image-based emotion recognition, multimodal and context-aware emotion categorization models have been presented recently. To name a few, a CAER-NET algorithm illustrated in the form of two stream encoding network which includes facial-stream and context-stream was proposed by [86]. In a similar way, authors in [88] present a multi-modal and context-aware emotion recognition framework that combines three interpretations of context from psychology principles. The first context uses multiple modalities of faces and gaits, the second encodes semantic information using self-attention-based CNN and the last one concerns socio-dynamic inter-agent interactions through depth maps. Regardless the current progress, still emotion categorization remains an open challenge for the artificial intelligence community.

4.2.2 Count data models

Existing methods for the problem of modeling count data have been based essentially on count response models including traditional methods like Poisson distribution, hurdle models, zero-inflated distribution, Negative Binomial, and multinomial response models. The multinomial distribution was usually the first candidate employed for modeling count data [64, 21, 10]. However, under Bayes assumption of independency, mutlinomial distribution fails to model the burstiness of words in documents. To this end, various alternatives have been proposed to enhance clustering performance and address the burstiness issue. Dirichlet distribution has always been known as the convenient conjugate prior for the multinomial distribution thanks to its good modeling capabilities and its statistical properties. The Dirichlet-compound-multinomial (DCM) which is the composition of the Dirichlet and the multinomial [22], has shown to be competitive with the best-known text classification methods by handling the word burstiness problem. Yet, it requires computationally expensive iteration method for learning its parameter. In addition, if the Dirichlet distribution is taken as a conjugate prior for DCM, the properties of such a prior have consequences on the regularities captured by the model. This Dirichlet prior relies on the facts that features with the same mean must have the same covariance and it has a very restrictive negative covariance structure [73]. In other side, taking into account that Dirichlet distribution has never been used for text generation and has been always considered as a prior such as in the case of DCM and Latent-Dirichlet allocation model, the smoothed Dirichlet [82] was the first text generator distribution proposed for modeling count data. This distribution is based on smoothing the count vectors to deal with the problem of sparseness and at the same time defining proportions in a smoothed subset of the whole simplex. Additionally, estimating the parameters with a maximum-likelihood approach presents a closed-form solution which makes the model simpler and faster than DCM. However, in the light of the limitations of Dirichlet distribution already mentioned, the smoothed Dirichlet suffers from the same disadvantages which motivates us to improve the distribution in order to address the challenges related to count data modeling.

Generalized Dirichlet distribution that has more general properties and robustness shows good capabilities for describing proportional data but has never been applied directly for modeling count data other than introduced as conjugate prior for the multinomial distribution [123]. This distribution arises in various contexts including: Bayesian life-testing problems [73, 55], mixture models for pattern recognition [56], and machine learning for computer vision [57]. When considering the generalized Dirichlet distribution, count vectors span the entire simplex. Yet, most vectors contain only a small fraction of the entire vocabulary which results in a high probability of zeros components corresponding to non-occupied data in the whole simplex.

In line with this view, we are proposing a novel count data model considered at first place as a text generator distribution defined in a smoothed simplex for emotion recognition. We are addressing several challenges of count data modeling through the following contributions:

(1) Provide an approach to tackle the sparseness and overdispersion problems with Smoothed Generalized Dirichlet (SGD) distribution.

(2) Propose a learning method for estimating the parameters and two clustering mechanisms namely mixture model based on EM algorithm and geometrical information using Kulback-Leibler, Fisher information, and Bhattacharyya distance.

(3) Present a new smoothed prior to the multinomial distribution to deal with burstiness and overdispersion problems; the Smoothed Generalized Dirichlet multinomial (SGDM); and a Newton-Raphson algorithm for learning the resulting model.

(4) Approximate the SGDM for high-dimensional count data using Taylor series expansion to present a new approximated distribution; Taylor approximation to the SGDM (TSGDM).

(5) Detect emotional states through two challenging applications. The first one addresses the problems of sparseness and burstiness in short texts that took place in social media "tweets" expressing the reaction of people due to disaster damages. The second application considers human pain intensity expressed through images and videos where we deal with the difficulties of high-dimensional count data and the burstiness of visual words.

4.3 Preliminary definitions and properties

In this section, we briefly review the count data properties, smoothed Dirichlet and the Generalized Dirichlet distributions which are necessary for building the proposed approaches.

Count data reflect the frequency of an event or the occurrence of words in text documents which lead to non-negative integer values or zeros. In the context of text, count data are represented using a vocabulary of documents where each document is described by an occurrence vector $X_i = (x_{i1}, ..., x_{id}, \ldots, x_{iD})$, where x_{id} denotes the number of times a word d appears in the document i. This type of data presents a considerable challenge for researchers as it is positively skewed with a high frequency of zeros values which induce the problem of sparsity. Besides, in a bag-of-words structure, the Naive Bayes assumption assumes that word emissions given a document are independent. Under this assumption, all the words are independent in the same document. This results in simpler, faster, and easier way to implement models but believing the word emissions are independent, where in fact is not, entails the burstiness phenomenon. In fact, burstiness appears as an accidental repetition of infrequent words in long documents where if a word has already appeared in a document, for instance, there is a higher probability of appearing again. The appearance of burstiness is not only revealed in words, but also propagated to the topics and count data in general. Count data do not only suffer from the sparsity and burstiness problem but also from high-dimensionality when considering large-scale dataset and overdispersion.

4.3.1 Smoothed Dirichlet

In the problem of count data modeling, the probability of generating a vector (e.g. a text document) that has a count representation is given mathematically as the product of the probabilities of the words. For this reason, it becomes common to consider multinomial-based models such as the DCM. A different generative process for count vectors is presented in [89, 82] where authors proposed a smoothed Dirichlet distribution based on generating a smoothing document model then unsmoothing it to get proportions X^u as follows:

$$X^u = (X^s - (1 - \lambda)X^{GE})/\lambda \tag{63}$$

where X^s is the smoothed proportions vector, X^{GE} is the proportions of words estimated from the entire document [82] and $0 < \lambda < 1$ is a smoothing parameter.

Taking into account that smoothed Dirichlet distribution has the same parametric form as the Dirichlet distribution but defining the vectors only in a compressed simplex, the probability density function of a smoothed D-dimensional vector is given as follows:

$$p(X^s|\vec{\alpha}) = \frac{1}{Z^{SD}(\vec{\alpha})} \prod_{d=1}^{D} (x_d^s)^{\alpha_d - 1} \tag{64}$$

where $\vec{\alpha}$ is the parameter of the smoothed Dirichlet and Z^{SD} is the smoothed Dirichlet normalizer that guarantees that the probabilities add up to 1. Considering the entire simplex $\Delta = \{X|\forall_d x_d > 0; \sum_{d=1}^{D} x_d = 1\}$, the normalizer should be:

$$Z^{SD} = \int_{\Delta} \prod_{d=1}^{D} (x_d^s)^{\alpha_d - 1} dX \tag{65}$$

However, given the smoothed vectors representation, the compressed simplex is given by the following:

$$\Delta^s = \{X^s\} = \{\lambda X^u + (1 - \lambda) X^{GE} | X^u \in \Delta\} \tag{66}$$

Thus, the smoothed Dirichlet distribution is expressed as follows:

$$p(X^s|\vec{\alpha}) = \frac{\prod_{d=1}^{D} (x_d^s)^{\alpha_d - 1} dX}{\lambda \int_{\Delta^s} \prod_{d=1}^{D} \{\lambda X^u + (1 - \lambda) X^{GE}\}^{\alpha_d - 1} dX^u} \tag{67}$$

For the purpose of simplifying the smoothed Dirichlet normalizer, the Gamma function is approximated using the Stirling's approximation that makes it unbounded and provides an approximation for the smoothed Dirichlet distribution [82] given by:

$$p(X^s|\vec{\alpha}) = \frac{\sum_{d=1}^{D} \alpha_d^{\sum_{d=1}^{D} \alpha_d}}{\prod_{d=1}^{D} \alpha_d^{\alpha_d}} \prod_{d=1}^{D} (x_d^s)^{\alpha_d - 1} \tag{68}$$

4.3.2 Generalized Dirichlet distribution

In Bayesian analysis, various types of generalized Dirichlet were proposed [124, 125, 28, 125, 74, 55] in the context of generalizing the Dirichlet distribution. All the common generalized Dirichlet distributions are derived based on the joint density function of $X = (x_1, \ldots, x_D) \in \Delta$ given by [124]:

$$p(X|\vec{\alpha}, \vec{\gamma}; \Lambda) = \frac{1}{C_{gD}} \Big(\prod_{d=1}^{D} x_d^{\alpha_d - 1} \Big) \prod_{q=1}^{Q} \Big(\sum_{d=1}^{D} \delta_{qd} x_d \Big)^{\gamma_q} \tag{69}$$

where C_{gD} is the normalizing constant, $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)$ is a positive parameter vector, $\vec{\gamma} = (\gamma_1, \ldots, \gamma_q)$ is a non-negative parameter vector, and $\Lambda = (\delta_{qd})$ is a $Q \times D$ matrix with $\delta_{id} = 0$ or 1 if there exists at least one non-zero element in each column of Λ.

As the normalizing constant has no explicit expression, many variant cases have been derived for proposing a convenient expression. In this paper, we consider the distribution of Connor and Mosimann [28] that is based on the concept of neutrality and offers a large structure for the covariance where variables with same mean do not have the same covariance and the covariance between two variables is not strictly negative as the case for Dirichlet distribution which ensures to face the problem of overdispersion. Assuming a stochastic representation where for each random vector $\vec{X}$, $z_1 = x_1, z_d = x_d/1 - x_1 - \cdots - x_d$ $(d = 2, \ldots, D)$, each z_d has a Beta distribution with parameters α_d and β_d. Thus, in a full simplex Δ, the normalizing constant is defined as follows:

$$\begin{aligned} C_{gD} &= \int_{\Delta} \prod_{d=1}^{D-1} (x_d)^{\alpha_d - 1} \big(1 - \sum_{k=1}^{d} x_k\big)^{\gamma_d} d\vec{X} \\ &= \prod_{d=1}^{D-1} \frac{\Gamma(\alpha_d)\Gamma(\beta_d)}{\Gamma(\alpha_d + \beta_d)} \end{aligned} \tag{70}$$

where $\gamma_d = \beta_d - \alpha_{d+1} - \beta_{d+1}, d = 1, \ldots, D - 2$ and $\gamma_{D-1} = \beta_{D-1} - 1$

4.4 Smoothed text generator distribution

The new distribution, SGD is obtained by smoothing the generalized Dirichlet and can be considered in the light of generalizing the smoothed Dirichlet distribution. We introduce first the new representation of count data vectors through smoothing functioning. Then, we provide the complete explanation of the new SGD distribution.

4.4.1 Smoothed generating proportion vectors

The major challenging problem of count data in text modeling is the sparsity where the unseen words are represented with zero probabilities. Supposing that we have a bag-of-words representation with a D vocabulary size resulting in count vectors $\vec{X} = (x_1, \dots, x_D)$, we first build the proportion vectors $\vec{P} = (p_1, \dots, p_D)$ normalized by the document length L. Then, to avoid assigning zero-probabilities to any word, we consider Jelinek-Mercer smoothing as considered in [81]. We smooth the proportion vectors by providing a smoothed parameter λ. The new smoothed vectors are generated as follows:

$$x_d^s = \lambda p_d + (1-\lambda) p_d^G, \ d = 1, \dots, D \tag{71}$$

where p_d^G is the general proportion that estimates how likely a word occurs in the entire set of documents. Given a set of N documents $(\vec{X}_1, \dots, \vec{X}_N)$, the p_d^G is defined as shown below:

$$p_d^G = \frac{\sum_{i=1}^N x_{id} + \delta}{\sum_{i=1}^N \sum_{d=1}^D x_{id} + D\delta} \tag{72}$$

where δ is a free parameter.

4.4.2 Smoothed Generalized Dirichlet and normalization

We assume that the probability of generating count vectors $\vec{X}$ under the proposed SGD distribution is the same as the probability of generating the smoothed proportion vectors $\vec{X}^s$ given in the following way:

$$\begin{aligned} P(\vec{X}^s|\vec{\alpha}, \vec{\beta}) &= \frac{1}{C_{SGD}(\vec{\alpha}, \vec{\beta})} \\ &\times \prod_{d=1}^{D-1} (x_d^s)^{\alpha_d - 1} \big(1 - \sum_{k=1}^{d} x_k^s\big)^{\gamma_d} \end{aligned} \tag{73}$$

where $C_{SGD}(\vec{\alpha}, \vec{\beta})$ is the SGD normalizing constant that guarantees the proportion probabilities to add up to 1, $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{D-1})$ and $\vec{\beta} = (\beta_1, \ldots, \beta_{D-1})$ are the parameters of the Smoothed Generalized Dirichlet that characterize the shape of the distribution, and $\gamma_d = \beta_d - \alpha_{d+1} - \beta_{d+1},\ d = 1, \ldots, D-2$ and $\gamma_{D-1} = \beta_{D-1} - 1$.

Exploiting the definition of a normalizing constant for the generalized Dirichlet distribution determined in a closed simplex Δ, we specify the SGD normalizing constant in the new compressed domain Δ^s by:

$$\begin{aligned} C_{SGD}(\vec{\alpha}, \vec{\beta}) &= \int_{\Delta^s} \prod_{d=1}^{D-1} (x_d^s)^{\alpha_d - 1} \big(1 - \sum_{k=1}^{d} x_k^s\big)^{\gamma_d} d\vec{X}^s \\ &= \int_{\Delta} \prod_{d=1}^{D-1} \{\lambda p_d + (1-\lambda) p_d^G\}^{\alpha_d - 1} \\ &\quad \big(1 - \sum_{k=1}^{d} \{\lambda p_k + (1-\lambda) p_k^G\}\big)^{\gamma_d} \lambda d\vec{P} \end{aligned} \tag{74}$$

As the normalizing constant does not have an explicit expression, we need to find an appropriate approximation defined in the new smoothed simplex. Inspired by the normalization of generalized Dirichlet described in section 4.3.2, the approximation of C_{SGD} depends mainly on Gamma function. Besides, in accordance with the smoothed Dirichlet distribution, authors in [82] suggested an approximation of Gamma function which provides an analytically tractable solution for the normalizing constant. Considering that Stirling's approximation of the Gamma function is the responsible of the unboundedness of the Dirichlet normalizer at smaller values ($\alpha \to 0$) when $\Gamma(\alpha) \to \infty$ as explained in [82], an approximation is defined based on disregarding the terms that lead to the infinity. Therefore, the approximated Gamma function is given by:

$$\Gamma_a(\alpha) \approx e^{-\alpha} \alpha^{\alpha} \tag{75}$$

where Γ_a is the approximation of Gamma function.

This approximated Gamma function gives us the opportunity to define the normalizing constant and to have finite values at the boundaries when $\alpha \to 0$ but closely identical to the exact function when

$\alpha \to \infty$. Then, the SGD-normalizer can be approximated by:

$$\begin{aligned} C_{SGD}(\vec{\alpha}, \vec{\beta}) &= \prod_{d=1}^{D-1} \frac{\Gamma_a(\alpha_d)\Gamma_a(\beta_d)}{\Gamma_a(\alpha_d + \beta_d)} \\ &= \prod_{d=1}^{D-1} \frac{e^{-\alpha_d}\alpha_d^{\alpha_d} e^{-\beta_d}\beta_d^{\beta_d}}{e^{-(\alpha_d+\beta_d)}(\alpha_d + \beta_d)^{\alpha_d+\beta_d}} \\ &= \prod_{d=1}^{D-1} \frac{\alpha_d^{\alpha_d}\beta_d^{\beta_d}}{(\alpha_d + \beta_d)^{\alpha_d+\beta_d}} \end{aligned} \tag{76}$$

Note that compared to the smoothed Dirichlet normalizer, the SGD normalizing constant has one extra parameter, which makes the distribution more flexible and gives rise to define the new Smoothed Generalized Dirichlet distribution as follows:

$$\begin{aligned} p(\vec{X}^s|\vec{\alpha}, \vec{\beta}) &= \prod_{d=1}^{D-1} \frac{(\alpha_d + \beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}} \\ &\quad (x_d^s)^{\alpha_d - 1}\big(1 - \sum_{k=1}^{d} x_k^s\big)^{\gamma_d} \end{aligned} \tag{77}$$

It is noteworthy to mention that SGD is reduced to smoothed Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_{D-1}, \alpha_D = \beta_{D-1})$ when $\beta_d = \alpha_{d+1} + \beta_{d+1}, \;\; d = 1, D-2$.

4.5 SGD parameters estimation and clustering mechanisms

We inaugurate this section by introducing the maximum likelihood approach proposed to estimate the SGD parameters. In the following, we present two clustering techniques: the mixture of SGD distributions using Bayes' rule and geometrical distances based agglomerative clustering.

4.5.1 Maximum Likelihood Estimation (MLE)

Considering an observed set of N smoothed vectors $\mathcal{X} = \{\vec{X}_i^s\}_{i=1}^N$, the observed-data log-likelihood function is given by:

$$\mathcal{L}(\mathcal{X}|\vec{\alpha}, \vec{\beta}) = \log \prod_{i=1}^{N} p(\vec{X}_i^s|\vec{\alpha}, \vec{\beta}) \tag{78}$$

When maximizing the log-likelihood function with respect to α_d and β_d parameters, we obtain a closed-form solution for each of them expressed as follows (the partial derivatives with respect to α_d and β_d are in Appendix A):

$$\hat{\alpha}_d = \sum_{i=1}^{N} \beta_d \Big(\frac{x_{id}^s}{1 - x_{id}^s} \Big) \quad (79)$$

$$\hat{\beta}_d = \sum_{i=1}^{N} \alpha_d \frac{\big(1 - \sum_{k=1}^{d} x_{ik}^s\big)}{\sum_{k=1}^{d} x_{ik}^s} \quad (80)$$

4.5.2 Mixture models based clustering

Even though, each and every distribution can have tremendous properties to deal with data modeling, a single one can only deal with unimodal data. Therefore, a natural choice to handle multimodality in real-world data applications is to take advantage of the combination of two or more distributions [126, 127]. For that matter, in our work, we consider a mixture of K SGD subcomponents. We propose the use of one of the standard likelihood-based estimation techniques for learning the SGD mixture models parameters: Expectation-Maximization (EM) algorithm [128, 47], [129], [130]. Beginning with a tuned initialization for the set of parameters, after the posteriors are inferred (named often as “responsibilities”) in the expectation step, then the iteration is proceeded to update the required variables until some convergence criterion is satisfied. The responsibilities or the posterior probabilities play an important role in likelihood-based estimation technique as they affect the update of the parameters in the next following step using the current parameter value. Then, by using the optimal mixture parameters, finding the best cluster for a given document is decided using Bayes' rule by choosing the cluster C corresponding to the document D whose parameters maximize the probability $p(C|D, \vec{\alpha}_c, \vec{\beta}_c)$ given by:

$$\begin{aligned} C_{best} &= \text{argmax}_C \log p(C|D, \vec{\alpha}_c, \vec{\beta}_c) \quad (81) \\ &= \text{argmax}_C \log p(D|\vec{\alpha}_c, \vec{\beta}_c, C)\pi_c \end{aligned}$$

where $\{\pi_c\}_{c=1}^{K}$ are the mixing weight coefficients subject to the constraints $\sum_{c=1}^{K} \pi_c = 1$ and $0 \leq \pi_c \leq 1$, and $p(D|\vec{\alpha}_c, \vec{\beta}_c, C)$ is the c-th probability density function (pdf) of the Smoothed Generalized Dirichlet distribution that corresponds to the cluster C. The complete learning algorithm is summarized in algorithm 3.

Algorithm 3: Smoothed Generalized Dirichlet mixture model learning algorithm

```
Input: $\mathcal{X}$ dataset, Number of components $K$ ;
Outpu: $\vec{\alpha}, \vec{\beta}$ ;
Initialization using K-means and the Method of Moment as in [123] ;
repeat
    foreach Component $c$ do
        Estimate the posterior distribution $r(i, c)$ using Bayes rule ;
        Estimate the mixing weight components $\pi_c = \frac{1}{N} \sum_{i=1}^{N} r(i, c)$ ;
        Update the parameters $\alpha_{cd}, \beta_{cd}$ $(d = 1, \ldots, D)$ using equations 79, and 80 ;
    end
until Convergence of Likelihood;
```

4.5.3 Geometrical information based clustering

Agglomerative clustering algorithms that use geometrical information distances are commonly important for processing different types of data [131, 132, 133]. Agglomerative methods are bottom-up hierarchical clustering approaches that build on merging pairwise clusters based on measuring the similarity or dissimilarity between groups. Measuring similarity is generally performed by means of Euclidean distance between data points. However, the dissimilarity measure is ususally performed using geometrical information such as Kulback-Leibler divergence [134], Bhattacharyya distance [135], Hellinger distance [136], Fisher information metric [137], Wasserstein distance [138] and many more. In this work, we propose an agglomerative clustering algorithm based on Kulback-Leibler divergence, Fisher information metric, and Bhattacharyya distance determined between Smoothed Generalized Dirichlet distributions. It is noteworthy to mention that Smoothed Generalized Dirichlet belongs to the family of exponential distributions which can be expressed in the following form:

$$p(\vec{X}^s|\vec{\alpha}, \vec{\beta}) = \exp(G(\Theta)^T T(X) + F(\Theta) + k(X)) \tag{82}$$

where

$$T(X) = [\log x_1^s, \dots, \log x_{D-1}^s - \log(1 - \sum_{k=1}^{D-2} x_k^s), \log(1 - x_1^s), \dots, \log(1 - \sum_{k=1}^{D-1} x_k^s) - \log(1 - \sum_{k=1}^{D-2} x_k^s)]$$

is the sufficient statistic,

$G(\Theta) = [\alpha_1, \dots, \alpha_{D-1}, \beta_1, \dots, \beta_{D-1}]$ are the natural parameters

$F(\Theta) = \sum_{d=1}^{D-1} (\alpha_d + \beta_d) \log(\alpha_d + \beta_d) - \alpha_d \log \alpha_d - \beta_d \log \beta_d$ is the log-normalizer,

$k(X) = 1$ is the carrier measure.

By considering the fact that Kulback-Leibler divergence between two distributions from exponential family is equivalent to the Bregman divergence, we express the following KL divergence between two SGDs as (Appendix A):

$$\begin{aligned}
\mathcal{K}(SGD_1 \parallel SGD_2) &= B_F(\vec{\alpha}_1, \vec{\beta}_1 \parallel \vec{\alpha}_2, \vec{\beta}_2) \\
&= H_F^{\times}(\vec{\alpha}_2, \vec{\beta}_2 \parallel \vec{\alpha}_1, \vec{\beta}_1) - H_F(\vec{\alpha}_2, \vec{\beta}_2) \\
&= \sum_{d=1}^{D-1} (\alpha_{d1} + \beta_{d1}) \log(\alpha_{d1} + \beta_{d1}) \\
&- (\alpha_{d2} + \beta_{d2}) \log(\alpha_{d2} + \beta_{d2}) \\
&- \alpha_{d1} \log \alpha_{d1} + \alpha_{d2} \log \alpha_{d2} \\
&- \beta_{d1} \log \beta_{d1} + \beta_{d2} \log \beta_{d2} \\
&+ \sum_{d=1}^{D-1} [\alpha_{d1} - \alpha_{d2}] \log \frac{\alpha_{d1}}{\alpha_{d1} + \beta_{d1}} \\
&+ [\beta_{d1} - \beta_{d2}] \log \frac{\beta_{d1}}{\alpha_{d1} + \beta_{d1}}
\end{aligned} \tag{83}$$

where H_F denotes the generalized entropy and $H_F^{\times}$ the generalized cross-entropy.

In the space of exponential families with two very close points, the Fisher metric is related to

KL-divergence by [139]:

$$\mathcal{F}(SGD_1 \parallel SGD_2) = \sqrt{2}\,\mathcal{K}(SGD_1 \parallel SGD_2)^{\frac{1}{2}} \tag{84}$$

As an additional probabilistic distance for which we take advantage of the properties of exponential families in the case of SGD, we develop a closed form expression for the Bhattacharyya distance using the following properties [140] (see Appendix A):

$$\mathcal{B}_h = \frac{1}{2}F(\Theta_1) + \frac{1}{2}F(\Theta_2) - F(\frac{1}{2}\Theta_1 + \frac{1}{2}\Theta_2) \tag{85}$$

where $F(\Theta_1)$ is the log-normalizer of SGD_1 and $F(\Theta_2)$ is the log-normalizer of SGD_2.

The agglomerative clustering algorithm starts with a large number of clusters. First, each object is represented by one cluster and we calculate the considered distance (KL, Fisher, Bhattacharyya) between each pairwise of SGD distributions. Next, clusters with the minimum distance are merged and the process is repeated until we found the target number of clusters.

4.6 Smoothed multinomial prior

Multinomial distribution has demonstrated capability to analyse count data in several applications [21, 10, 64]. However, a central problem with this distribution is the so-called "Naïve Bayes assumption". This hypothesis imposes the independence of all the attributes and consequently words are generated separately. The multinomial independency assumption hinders capturing the phenomenon of burstiness where a word that has already appeared in a document, for instance, has a higher probability of appearing again [4, 3]. To overcome the burstiness deficiency, DCM, a distribution that combines the multinomial and the Dirichlet as a prior has been introduced to smooth the multinomial parameters. In fact, the DCM model has achieved good results in various applications [141, 142, 11, 26, 143, 25, 24, 23]. However, the Dirichlet distribution has a restrictive covariance matrix which makes this model impractical in real-world applications. Addressing this disadvantage, a number of priors which belong to the family of generalized Liouville distributions [102] have been considered as alternative. Taking into account that Smoothed Generalized Dirichlet is also a conjugate prior to the multinomial distribution (see Appendix A), it is interesting to propose a new smoothed multinomial prior to develop a new count data modeling approach. Consider the multinomial distribution, we are modeling the probability of count variables appearing in one of D possible emissions, $P_1, \ldots, P_D$. When we consider the bag-of-words structure, each document is represented by the set of word occurrences and drawn from a multinomial distribution of words. Let $\vec{X}$ be a vector of word counts which follows a multinomial distribution with parameters $\vec{P} = (P_1, \ldots, P_D)$:

$$p(\vec{X}|\vec{P}) = \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D} P_d^{x_d} \tag{86}$$

where $|x| = \sum_{d=1}^{D} x_d$, the parameter P_d is the probability of emitting a word d from the document represented by $\vec{X}$, $0 \leq P_d \leq 1$ and $\sum_{d=1}^{D} P_d = 1$.

Our aim is to express each document over a count-data vector $\vec{X}$ by a multinomial distribution whose parameters $\vec{P}$ are generated by the novel Smoothed Generalized Dirichlet distribution. In this case, we determine the joint distribution of the vector $\vec{X}$ and $\vec{P}$ (see Appendix A). Thus, marginalizing out $\vec{P}$ vector for the multinomial weighted by the Smoothed Generalized Dirichlet distribution gives us the likelihood of a document using the SGDM approach:

$$\begin{aligned}
p(\vec{X}|\vec{\alpha},\vec{\beta}) &= \int_{\vec{P}} p(\vec{X}|\vec{P})p(\vec{P}|\vec{\alpha},\vec{\beta})d\vec{P} \qquad (87)\\
&= \int_{\vec{P}} \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D} P_d^{x_d} \prod_{d=1}^{D-1} \frac{(\alpha_d+\beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}} (P_d^s)^{\alpha_d-1}\big(1-\sum_{k=1}^{d} P_k^s\big)^{\gamma_d} d\vec{P} \\
&= \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D-1} \frac{(\alpha_d+\beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}} \int_{\vec{P^s}} \prod_{d=1}^{D} (P_d^s)^{x_d}(P_d^s)^{\alpha_d-1}\big(1-\sum_{k=1}^{d} P_k^s\big)^{\gamma_d} d\vec{P^s} \\
&= \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D-1} \frac{(\alpha_d+\beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}} \int_{\vec{P^s}} \prod_{d=1}^{D} (P_d^s)^{\alpha'_d-1}\big(1-\sum_{k=1}^{d} P_k^s\big)^{\gamma'_d} d\vec{P^s} \\
&= \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D-1} \frac{(\alpha_d+\beta_d)^{\alpha_d+\beta_d}(\alpha'_d)^{\alpha'_d}(\beta'_d)^{\beta'_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}(\alpha'_d+\beta'_d)^{\alpha'_d+\beta'_d}}
\end{aligned}$$

where $\alpha'_d = \alpha_d + x_d$ and $\beta'_d = \beta_d + \sum_{k=d+1}^{D} x_k$ for $d = 1, \ldots, D-1$, $\gamma'_d = \beta'_d - \alpha'_{d+1} - \beta'_{d+1}$

for $d = 1, \dots, d-2$ and $\gamma'_{D-1} = \beta'_{D-1} - 1$.

For estimating the $\vec{\alpha}, \vec{\beta}$ parameters of SGDM model, we consider a maximum likelihood approach. Though, no closed-form solution exists. We therefore estimate these parameters in accordance with the Newton-Raphson method:

$$\Theta^{(t+1)} = \Theta^{(t)} - H(\Theta^{(t)})^{-1} \frac{\partial L(\mathcal{X}|\Theta^{(t)})}{\partial \Theta^{(t)}} \tag{88}$$

where $\Theta = (\vec{\alpha}, \vec{\beta}), H(\Theta^{(t)})^{-1}$ is the inverse of the Hessian matrix and $L(\mathcal{X}|\Theta^{(t)})$ is the log-likelihood of the SGDM distribution. To make the Newton-Raphson method, we straightforwardly compute the first, second, and mixed derivatives of $L(\mathcal{X}|\Theta^{(t)})$ with respect to $\vec{\alpha}$ and $\vec{\beta}$.

By computing the first derivatives, we obtain the following, for $d = 1, \dots, D-1$

$$\begin{aligned} \frac{\partial L(\mathcal{X}|\Theta^{(t)})}{\partial \alpha_d} &= \{\log(\alpha_d + \beta_d) - \log \alpha_d\} \\ &+ \sum_{i=1}^{N} \{\log(\alpha_d + x_{id}) \\ &- \log(\alpha_d + \beta_d + \sum_{k=d}^{D} x_{ik})\} \end{aligned} \tag{89}$$

$$\begin{aligned} \frac{\partial L(\mathcal{X}|\Theta^{(t)})}{\partial \beta_d} &= \{\log(\alpha_d + \beta_d) - \log \beta_d\} \\ &+ \sum_{i=1}^{N} \{\log(\beta_d + \sum_{k=d+1}^{D} x_{ik}) \\ &- \log(\alpha_d + \beta_d + \sum_{k=d}^{D} x_{ik}))\} \end{aligned} \tag{90}$$

Subsequently, determining the Hessian matrix requires determining the second and mixed derivatives, where $H(\alpha_d, \beta_d)$ is defined for $d = 1, \dots, D-1$ as:

$$H(\alpha_d, \beta_d) = \begin{bmatrix} \frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial^2 \alpha_d} & \frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial \alpha_d \partial \beta_d} \\ \frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial \beta_d \partial \alpha_d} & \frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial^2 \beta_d} \end{bmatrix}, \tag{91}$$

where

$$\frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial \alpha_{d1} \partial \alpha_{d2}} = \begin{cases} \frac{1}{\alpha_d+\beta_d} - \frac{1}{\alpha_d} + \sum_{i=1}^{N} \frac{1}{\alpha_d+x_{id}} & \\ -\frac{1}{\alpha_d+\beta_d+\sum_{k=d}^{D} x_{ik}} & \text{if } d_1 = d_2 = d, \\ 0 & \text{Otherwise} \end{cases} \tag{92}$$

$$\frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial \beta_{d1} \partial \beta_{d2}} = \begin{cases} \frac{1}{\alpha_d+\beta_d} - \frac{1}{\beta_d} + \sum_{i=1}^{N} \frac{1}{\beta_d+\sum_{k=d+1}^{D} x_{ik}} & \\ -\frac{1}{\alpha_d+\beta_d+\sum_{k=d}^{D} x_{ik}} & \text{if } d_1 = d_2 = d, \\ 0 & \text{Otherwise} \end{cases} \tag{93}$$

$$\frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial \alpha_{d1} \partial \beta_{d2}} = \frac{\partial^2 L(\mathcal{X}|\Theta^{(t)})}{\partial \beta_{d1} \partial \alpha_{d2}} = \begin{cases} \frac{1}{\alpha_d+\beta_d} & \\ -\sum_{i=1}^{N} \frac{1}{\alpha_d+\beta_d+\sum_{k=d}^{D} x_{ik}} & \text{if } d_1 = d_2 = d, \\ 0 & \text{Otherwise} \end{cases} \tag{94}$$

Following, we need to determine the inverse of the Hessian matrix that can be simplified as (Theorem 8.3.3 in [144]):

$$H(\alpha_d, \beta_d) = D + \delta a a^{tr} \tag{95}$$

where

$$D = \text{diag}\Big[-\frac{1}{\alpha_d} + \sum_{i=1}^{N} \frac{1}{\alpha_d + x_{id}}, -\frac{1}{\beta_d} + \sum_{i=1}^{N} \frac{1}{\beta_d + \sum_{k=d+1}^{D} x_{ik}} \Big] \tag{96}$$

$\delta = \frac{1}{\alpha_d+\beta_d} - \sum_{i=1}^{N} \frac{1}{\alpha_d+\beta_d+\sum_{k=d}^{D} x_{ik}}$, and $a^{tr} = 1$.

Then, the inverse of the matrix is defined as:

$$H(\alpha_d, \beta_d)^{-1} = D^* + \delta^* a^* a^{*tr} \tag{97}$$

where

$$\begin{aligned} D^* &= D^{-1} = \text{diag}[1/D_1, 1/D_2] \\ &= \text{diag}\left[\frac{1}{-\frac{1}{\alpha_d} + \sum_{i=1}^{N} \frac{1}{\alpha_d + x_{id}}}, \frac{1}{-\frac{1}{\beta_d} + \sum_{i=1}^{N} \frac{1}{\beta_d + \sum_{k=d+1}^{D} x_{ik}}}\right], \end{aligned} \tag{98}$$

$$\begin{aligned} \delta^* &= -\delta\Big(1 + \delta(1/D_1 + 1/D_2)\Big)^{-1} \\ &= \Big(\sum_{i=1}^{N} \frac{1}{\beta_d + \sum_{k=d}^{D} x_{ik}} - \frac{1}{\alpha_d + \beta_d}\Big) \\ &\left(1 + \frac{\frac{1}{\alpha_d+\beta_d} - \sum_{i=1}^{N} \frac{1}{\alpha_d+\beta_d+\sum_{k=d}^{D} x_{ik}}}{-\frac{1}{\alpha_d} + \sum_{i=1}^{N} \frac{1}{\alpha_d+x_{id}}} + \frac{\frac{1}{\alpha_d+\beta_d} - \sum_{i=1}^{N} \frac{1}{\alpha_d+\beta_d+\sum_{k=d}^{D} x_{ik}}}{-\frac{1}{\beta_d} + \sum_{i=1}^{N} \frac{1}{\beta_d+\sum_{k=d+1}^{D} x_{ik}}}\right)^{-1} \end{aligned} \tag{99}$$

and

$$\begin{aligned} a^{*tr} &= (a_1/D_1, a_2/D_2) \\ &= \left(\frac{1}{-\frac{1}{\alpha_d} + \sum_{i=1}^{N} \frac{1}{\alpha_d + x_{id}}}, \frac{1}{-\frac{1}{\beta_d} + \sum_{i=1}^{N} \frac{1}{\beta_d + \sum_{k=d+1}^{D} x_{ik}}}\right) \end{aligned} \tag{100}$$

4.7 Application: Disaster tweets related emotions

4.7.1 Experimental data and procedures

In this section, we evaluate the proposed approaches on three datasets: 2014 India floods, 2014 California earthquake, and 2013 Pakistan earthquake selected from the Crisis corpora collection [145]. The corpora collection is constructed with tweets messages related to 19 different crisis that took place in different countries between 2013 and 2015 such as floods, earthquake, typhoon, and infectious disease. The 2014 India floods includes 5,259,681 tweets, 2014 California earthquake dataset contains 254,525 tweets, and 156,905 tweets messages are related to 2013 Pakistan earthquake. The crisis-related messages are annotated by information types assigned to 9 categories as shown in Table 4.1 using a subset of the annotations employed by the United Nations Office for the

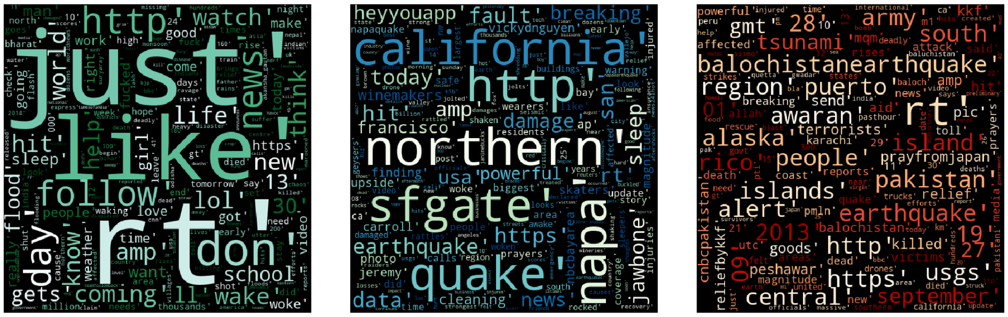


Figure 4.1: Wordcloud for three selected datasets: India floods, California eqarthquake, and Pakistan earthquake

Coordination of Humanitarian Affairs (UN OCHA). This corpora presents two important challenges as the datasets have high class imbalance and collected messages are introduced as short texts which increase the probability of zeros values. According to Figure 4.1, the considered datasets contain millions of words which are not all related to the crisis categories that continues to present a challenge for our modeling.

In our method, the bag-of-words approach is used to represent the tweets messages as count vectors. We represent each short text "tweets" as a D-dimensional count vector $(D = 10, 20, \ldots, 100)$. After, we smooth the vectors to adapt them to the proposed approaches. Then, we cluster each dataset using the proposed models: SGD-based mixture, SGD-based KL, SGD-based Fisher, SGD-based Bhattacharyya, and SGDM. We note that the clustering mechanism of the SGDM is based only on mixture model as the distribution does not belong to the exponential family and we are not able to take advantage of the properties of geometrical distances in this case.

Table 4.1: Description of the tweets emotion categories for the three selected datasets

Tweet Category	Description
Injured or dead people	Reports of casualties and/or injured people due to the crisis
Missing, trapped, or found people	Reports and/or questions about missing or found people
Displaced people and evacuations	People who have relocated due to the crisis, even for a short time (includes evacuations)
Infrastructure and utilities damage	Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored
Donation needs or offers or volunteering services	Reports of urgent needs or donations of shelter, and volunteering services
Caution and advice	Reports of warnings issued or lifted, guidance and tips
Sympathy and emotional support	Prayers, thoughts, and emotional support
Other useful information	Other useful information that helps understand the situation
Not related or irrelevant	Unrelated to the situation or irrelevant

4.7.2 Evaluation results

In order to evaluate the robustness of the proposed unsupervised approach, we compare our methods with the related count-data models namely: multinomial mixture, DCM mixture model, generalized Dirichlet multinomial (GDM) mixture, and smoothed Dirichlet distribution and the Gaussian mixture model (GMM). The experimental results of the proposed models and the related-work methods are presented in Table 4.2 in terms of accuracy and running time. We note that the machine used for the experimentation has the following characteristics (Processor: Intel i5 CPU @1.6 GHz, RAM: 8 GB). We observe that SGDM model achieves the best performance on California earthquake and Pakistan earthquake datasets and SGD mixture model outperforms the other methods for the India floods dataset. We note that the results of SGD based on Kulback-Leibler and Fisher metric are slightly better than Bhattacharyya distance. As expected, the smoothed approaches present better results when dealing with short texts that suffer from the sparseness problem. Besides, we clearly observe the speed up of the SGD mixtures over DCM and GDM. We did not mention the running time of the geometrical information clustering mechanism as it is unfair to compare them due to applying the modeling for each text separately while using the mixture-based, clustering is applied only once for the whole database. It is worth mentioning that SGD-based geometrical information outperforms the other methods but not as much as SGD-based mixture. Consider, for instance, the strength of mixture model clustering vs agglomerative clustering where the multimodality is taken into account in the first mechanism contrary to the second when only one distribution is applied.

Figure 4.2 illustrates the influence of smoothing parameter and the vocabulary size on the performance of Smoothed Generalized Dirichlet mixture-based clustering. We mention that the best performances on California earthquake and Pakistan earthquake are obtained using a vocabulary size of 10 features and a smoothing parameter λ equals to 0.6 and 0.2, respectively. Regarding the India floods, the best accuracy is achieved with 40-dimensional vectors length and λ = 0.1. It can be clearly observed that more the vocabulary size is lower more the smoothing parameter affects the results. Along with a large vocabulary, the results become identical for all the values of smoothing parameters. Besides, from Figure 4.3, we evaluate the performance of the Smoothed Generalized

Dirichlet multinomial in relation to the vocabulary size. We point out the influence of the dimension of count data on the performance. As we are considering in this application only short texts, we did not address the high-dimensionality issue for the proposed models and we only take into account features length between 10 and 100. In this matter, we observe that SGDM obtains best performance when using 40-dimensional features for India floods and California earthquake datasets and 10-dimensional vectors for Pakistan earthquake.

Table 4.2: Clustering results in terms of accuracy for selected datasets (Tweets messages) using different count-data models

Models	India floods (9 clusters)		California earthquake (9 clusters)		Pakistan earthquake (9 clusters)	
	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)
Gaussian mixtures	06.81	1.12	11.34	0.20	07.97	1.24
Multinomial mixtures	37.20	0.19	28.27	0.10	23.81	0.13
Dirichlet compound multinomial	34.61	505.26	23.69	206.31	28.76	559.05
Smoothed Dirichlet	50.65	27.81	31.09	14.89	38.49	16.94
Generalized Dirichlet multinomial	49.06	260	40.02	100.74	38.49	48.10
Smoothed Generalized Dirichlet-mixture	**71.31**	92.87	51.26	34.92	41.71	40.35
Smoothed Generalized Dirichlet- KL	56.64	-	49.20	-	40.34	-
Smoothed Generalized Dirichlet- Fisher	56.53	-	49.91	-	41.03	-
Smoothed Generalized Dirichlet- Bhatt	49.12	-	50.85	-	39.38	-
Smoothed Generalized Dirichlet multinomial	61.37	108.23	**53.32**	45.93	**47.42**	53.25

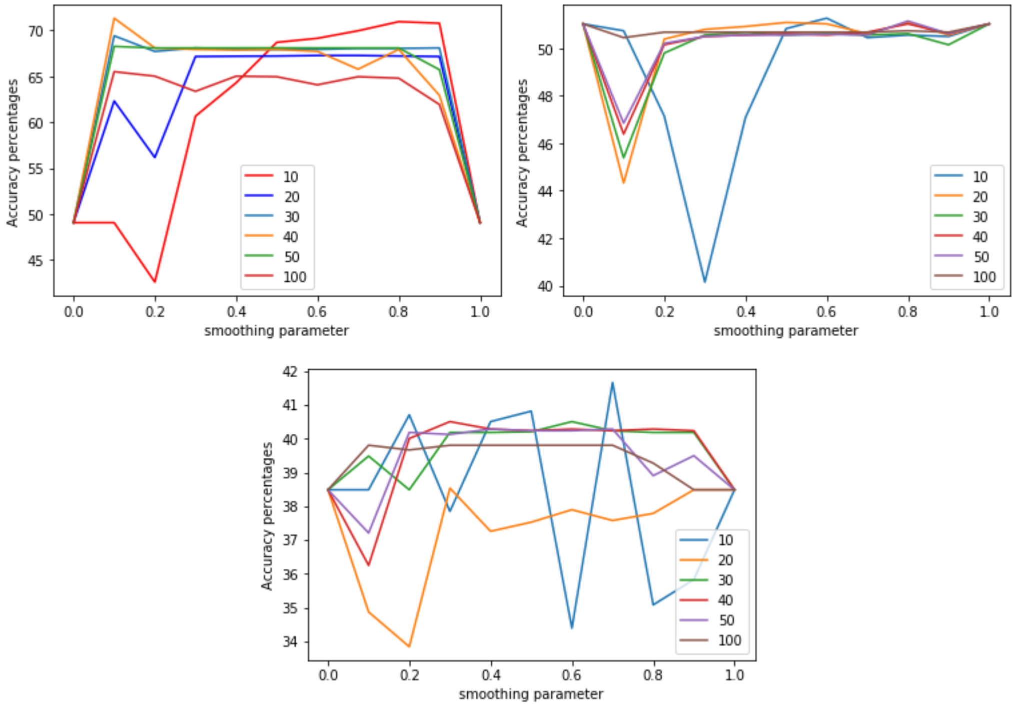


Figure 4.2: Evaluation of SGD mixure based clustering in terms of smoothing parameter and the size of vocabulary for three selected datasets; the first row from left to right: India floods, Californa earthquake, the second row: Pakistan earthquake

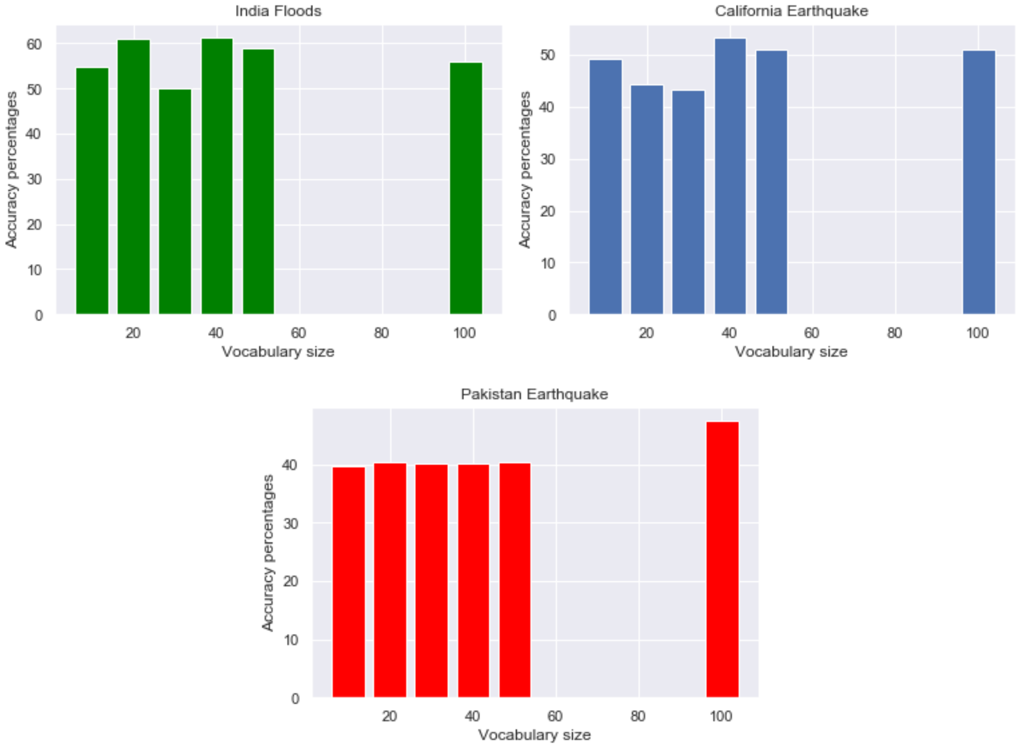


Figure 4.3: Evaluation of Smoothed generalized Dirichlet multinomial in terms of vocabulary size for three selected datasets

4.8 Application: Pain intensity estimation

4.8.1 Experimental data and procedures

In order to evaluate the robustness of the proposed approaches in real-life applications, we consider the estimation of pain intensity through video frames. We experiment the count data challenges through image features taking into account the burstiness problem and the high-dimensionality issue. We conduct extensive experiments on a publicly available database namely BioVid Heat Pain (BioVid) [100]. The BioVid database is collected within a study of 90 males and females participants from three age groups (18-35, 36-50, and 51-65). The database contains different types of data such as the skin conductance level (SCL), the electrocardiogram (ECG), the electroencephalogram (EEG) and videos sequences provided by Kinect. In our work, we consider Part A (facial video data) of this database which includes five pain intensity levels (see Figure 4.4): no pain (level 0 or baseline BN), low pain (level one, PA1), intermediate pain (levels 2 and 3, PA2 and PA3), and severe

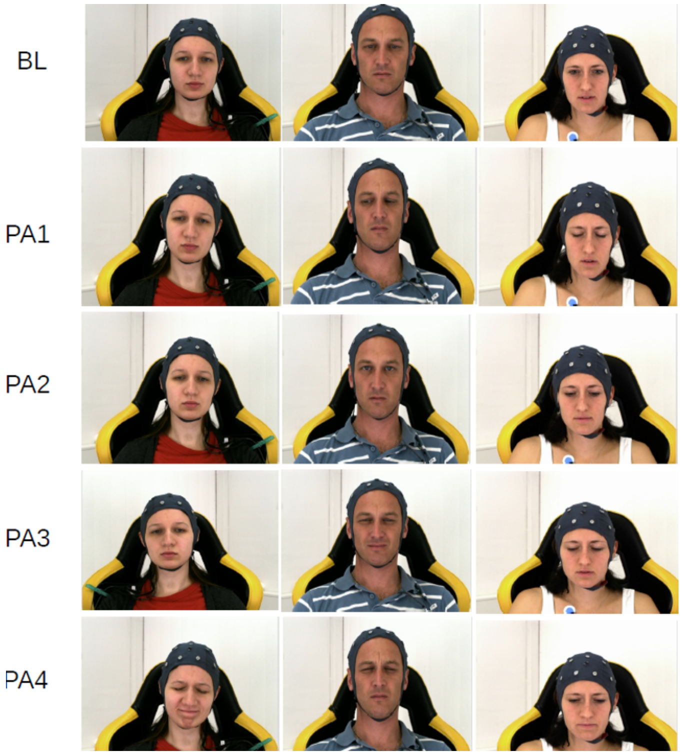


Figure 4.4: Face samples from Biovid-Heat Pain dataset accross pain intensities (PA*) and baseline (BL)

pain (level 4, PA4). The total number of videos is 8,700 sequences where there are 100 samples per person making 1,740 videos per intensity levels. This database contains challenging videos due to describing pain is dissimilar from one person to another. As it is shown in Figure 4.4, it is observed that three participants have different facial indicators of pain which make it difficult to detect pain and to estimate intensity of facial actions associated with pain. In our experiments, we conduct a pain detection scenario. We followed the protocol corresponding to Biovid database in the work of [101] which is designed on frame level (Pain vs No Pain). Thus, the frames recorded with no simulation (BL) considered as No Pain are compared to the ones with pain level 3 and 4 (PA3, PA4) taken as the Pain class. Hence, we use 20 subjects for each pain level and we extract frames from each video which results in 11,005 images (No Pain: 5,520 and Pain: 5,485). For the feature extraction, we employ LBP descriptor which has shown to be the most adequate descriptor for facial expressions and we construct from these features a bag-of-visual-words using different numbers of visual vocabulary sizes $(50, 100, \ldots, 2500)$ for the purpose of presenting the images/frames into count data. We study the performance on various number of frames starting from the last 20 frames in each video till the total number of frames extracted from a sequence.

4.8.2 Evaluation criteria

From Table 4.3, we remark that considering only 20 frames to represent a sequence is not enough and using the totality of frames is not optimizing in terms of efficiency and running time as the first seconds of the sequence could contains non important information regarding the pain facial expressions. Further, the best results are obtained using the last 80 frames that contain the most significant expressions. It is noteworthy to mention that an appropriate tuning of the initialization of $\vec{\alpha}$ and $\vec{\beta}$ parameters plays an important role on the performance of our proposed models. We mention also that using the visual words, the models have been also affected by vocabulary sizes and smoothing parameter as shown in disaster tweets related emotions application. It is clearly observed that our proposed smoothed approaches outperform the other multinomial-based methods which demonstrate the efficiency in addressing the challenges of visual words namely the sparseness, the overdispersion, and the burstiness. Among the multinomial models, our SGDM achieves the best accuracy as well as the SGD based on mixture and agglomerative clustering that accurately determines the pain facial expressions while the multinomial, the DCM, and the GDM are not appropriate for such challenging database. Further, Table 4.4 shows a comparison between the proposed approaches and state-of-the-art methods applied on the BioVid database for pain detection where we achieve superior results of 86.55% for detecting pain expressions using SGDM.

4.8.3 Taylor approximation to the SGDM

For high-dimensional count data, the multinomial-based models are usually suffering from high-computational problems due to the non-closed form solution for learning parameters. In the case of SGDM, taking into account the high-dimensional features, the model takes more time for estimating the parameters despite that it outperforms all the other models in terms of recognizing the emotional states for both text and images. For the sake of optimizing the model, we propose an approximation that is based on the fact that parameters $\vec{\alpha}$ and $\vec{\beta}$ are too small positive values and very close to zero in a high-dimensional feature space (details are in supplementary materials). The new approximated

Table 4.3: Accuracy results for Pain detection on BioVid-Heat Pain Dataset using different count-data models

Models	Number of Frames					
	20	40	60	80	100	138
Multinomial	55.56	55.88	56.57	52.42	53.06	51.46
DCM	55.22	54.87	56.14	53.83	53.06	52.83
GDM	49.89	49.97	50.36	50.28	50.22	50.16
SD	55.56	60.40	60.40	69.60	61.56	56.56
SGD- KL	53.97	71.05	72.60	58.66	60.83	60.32
SGD- Fisher	53.97	71.05	72.60	58.66	60.83	60.32
SGD- Bhatt	50.11	75.55	62.40	58.66	65.36	60.80
SGD mixtures	56.81	73.50	74.90	80.50	65.33	64.80
SGDM	52.84	75.22	78.00	86.55	62.20	75.40

Table 4.4: Comparative analysis on the BioVid database for Pain detection (Frame level)

Method	Accuracy (%)
LBP [101]	59.08
LPQ [101]	58.82
BSIF [101]	59.25
LBP + LPQ [101]	60.23
LPQ+BSIF [101]	59.83
LBP + BSIF [101]	60.23
GMM	22.36
SGD mixtures	80.50
SGD-KL	72.60
SGD-Fisher	72.60
SGD-Bhatt	75.55
SGDM	86.55

count-data distribution, so-called Taylor approximation to the SGDM (TSGDM), with prescribed parameters $\vec{\alpha}$, $\vec{\beta}$ is defined by:

$$\begin{aligned} p(\vec{X}|\vec{\alpha},\vec{\beta}) \quad \simeq \quad & \frac{|x|!}{\prod_{d=1}^{D} x_d!} \prod_{d=1}^{D} \frac{1+x-d\log(x_d)}{1+x_d\log(\beta-d+Z_d)} \\ & \frac{1+Z_{d+1}\log(Z_{d+1})}{1+Z_{d+1}\log(\alpha_d+Z_d)} \end{aligned} \tag{101}$$

We estimate the parameters α_d and β_d using MLE (Maximum-Likelihood Estimator) learning method where we obtain closed-form solutions which motivate our interest in optimizing the computation complexity. We provide the following log-likelihood for a set of count vectors $\vec{X}_1, \ldots, \vec{X}_N$:

$$\begin{aligned} \log\prod_{i=1}^{N} p(\vec{X}_i|\vec{\alpha},\vec{\beta}) \quad = \quad & \sum_{i=1}^{N} \Big(\log|x|! + \sum_{d=1}^{D} \log(1+x_d\log(x_d)) + \log(1+Z_{d+1d+1}) \qquad (102) \\ - \quad & \log(1+x_d\log(\beta_d+Z_d)) - \log(1+Z_{d+1}\log(\alpha_d+Z_d)) - \sum_{d=1}^{D} \log(x_d!) \Big) \end{aligned}$$

When evaluating the log-likelihood with respect to the parameters, the resulting inference gives the following:

$$\hat{\alpha}_d = \frac{Z_{d+1}}{1+Z_{d+1}\log(\alpha_d+Z_d)} - Z_d \tag{103}$$

$$\hat{\beta}_d = \frac{x_d}{1+x_d\log(\beta_d+x_d)} - x_d \tag{104}$$

We investigate the performance of the proposed TSGDM on the BioVid database. In this matter, we compare the running time and accuracy of the proposed models (SGD mixtures, SGDM) with TSGDM approach under various dimensions of LBP features in the case of 80 frames. It is clearly observed from Table 4.5, that SGD mixtures and SGDM performance declines when the size of vocabulary becomes larger. Despite the outstanding results of these two models, they are unable to support the high-dimensionality issue. We note that SGDM takes more time for running than the SGD mixtures due to the no-closed form solution. A significant improvement was seen between

Table 4.5: Accuracies results for Pain detection on BioVid-Heat Pain Dataset using different count-data models

Proposed Models	Features dimension									
	50		100		700		1100		2500	
	Accuracy	Time (min)	Accuracy	Time (min)	Accuracy	Time (min)	Accuracy	Time (min)	Accuracy	Time (min)
SGD mixtures	**80.50**	6.53	79.80	24.03	62.00	355.33	52.00	796.11	52.33	2,620.05
SGDM	**86.55**	14.31	57.01	52.9	68.00	640.95	53.30	1560.45	54.25	3,865.6
TSGDM	81.02	3.41	**84.30**	12.11	60.15	141.48	60.00	352.75	54.00	1,810.4

SGDM and the approximated distribution TSGDM in terms of running time and comparable results with regard to accuracy. Accordingly, the SGDM is the best model in terms of accuracy but when we compare runtime with regards to other proposed models, it tooks more time in case of high-dimensional data. For that, we proposed the approximation for more simplification and fastness. Thus, when we have data where the features don't exceed "700" size of vocabulary, the SGDM is the best choice to use it for modeling and if we have high-dimensional data, it is preferable to apply the TSGDM.

4.9 Conclusion

Count data modeling continues to attract considerable interest in data mining community. In this work, we proposed novel probabilistic approaches: SGD mixtures, SGD-KL, SGD-Fisher, SGD-Bhattacharyya, SGDM, and TSGDM. The core idea of our models is to deal with the problems of count data such as the sparseness, the overdispersion, the burstiness, and the high-dimensionality issue. In this regard, the count data is smoothed using Jelinek-Mercer smoothing approach to avoid assigning zero-probabilities and a new generalized Dirichlet distribution is defined on a smoothed simplex that is due to its covariance properties was able to face the overdispersion problem. Regarding parameters learning, we considered a maximum likelihood approach for the SGD and two different clustering mechanisms: mixture modeling and agglomerative-based geometrical information (Kulback-Leibler, Fisher information, Bhattacharyya distance).

Taking into account the burstiness phenomenon, we provided a new smoothed prior for the multinomial distribution to present a new count-data modeling so-called Smoothed Generalized Dirichlet multinomial. Further, dealing with high-dimensional features, we approximated the SGDM using Taylor series expansion to propose the TSGDM model.

For the sake of evaluating the proposed models, we applied them to detect emotional states. For that, we selected two different applications: disaster tweets related emotions and pain intensity estimation. Experimental results indicate the superior efficiency and robustness of our approaches in modeling text and images features. Hence, the proposed smoothed models outperform the competing count data models and methods reported in literature. Taking into account the impact of inference techniques on the performance of probabilistic models, we plan to investigate estimation techniques other than frequentest approaches such as variational learning that has achieved impressive results in non-linear model inference problems. In view of the outstanding obtained results, it is interesting to consider providing other smoothed multinomial based approaches as a future work as well as proving the utility of the proposed models in other complex applications.

Chapter 5

Latent Smoothed Beta-Liouville Topic Modeling for Emotion Analysis and Affect Recognition

This chapter is concerned with mining emotions from child-directed texts and recognizing human affect states from face and body expressions. In recent years, several research works have been addressed to multi-modal emotion recognition from facial, body, voice, and physiological signals. Emotions expressed by people from text as well may take multiple formats such as messages, tweets, letters, and books. This work explores new ways of emotion analysis from fairy tales and visual affect recognition from two modalities: face and body. First, we propose a new statistical approach based on Beta-Liouville distribution and a smoothed simplex for the purpose of addressing textual data challenges. Second, we extend the novel distribution to topic modeling where we model data (text/image) as a mixture of Smoothed Beta-Liouville (SBL) distributions which we represent by latent topics and we introduce a clustering algorithm through an Expectation-Maximization approach. Following, we incorporate modeling unknown documents in a Bayesian folding-in way where we present thereby a novel Emotion-Term model using SBL kernels for PLSI. Third, we apply our new models for tracking emotions in fairy tales where experiments results show the outperformance of the proposed models with regards to the related smoothed-based models that prove the aptitude of addressing textual data challenges. Next, we study the challenges of bimodal affect recognition from face and body through Latent SBL. Experiments on FABO database demonstrate that Latent SBL achieves better recognition rates than related-works and even outperforms supervised algorithms in some scenarios.

5.1 Introduction

Emotion is a psychological state defined as a reaction pattern of such experience involving behavioral elements. Humans perceive their emotions through different modalities such as facial expressions, body language, verbal signals, and textual context. Categorizing human emotions becomes essential both in developing efficient Artificial Intelligence (AI) systems and in understanding how humans convey their psychological states. Emotion categorization plays also a vital role in human-computer interaction. For instance, a number of robotics applications have became possible thanks to facial expression recognition. Not only robotics technologies have been invested in emotion AI but also education [146], healthcare [147], natural disaster prediction [17], and personalized recommendation [148]. These recent areas involve different modalities of emotion recognition to help industries improve customer satisfaction. For example, in education, thanks to emotion-based AI, new learning systems have been adapted to autistic children where they can recognize their level of frustration. As well, detecting stress and anxiety level are currently considered in numerous companies for employee safety.

Accordingly, multiple research studies have been directed to emotion recognition from human facial expression, speech signals, and text messages. While facial expression and emotional speech signals have been studied extensively, text-based emotion recognition received recently attention due to the rapid growth of AI and in particular Natural Language Processing (NLP) which combines computational and linguistic techniques to help computers understand human languages in the format of text. Extracting emotions from textual documents demonstrates massive challenges including the complexity of meaning, the ambiguity of words in the text, and writing styles. Understanding the verbal meaning was even a challenge for human to distinguish their own emotional states. Initially, AI models start with six fundamental emotional states (happiness, sadness, anger, disgust, surprise, and fear) which are defined by Ekman [149] and extended by Alm [150] in which he included positive surprise and negative surprise. Recently, larger datasets involve more emotional states such as "shame" suggested by Emotion-Stimulus database [151], "trust" and "anticipation" defined by Mohammad et al. [152], and Crowdflower [1] which introduces new categories consisting of thirteen emotions: fun, worry, love, hate, surprise, happiness, sadness, anger, enthusiasm, boredom, relief, and empty. For instance, emotion recognition (ER) models consider wider applications that consist of detecting mental states such as frustration [153], unsure, and concentrating [154].

More recently, human body gestures have been linked to emotions understanding. Early research works on gestures have been oriented basically to recognize human activities but investigating the relation between gesture and emotions expressions was sparsely covered. Previous works aim to detect behaviors using illustrative gestures such as waving hands to say "HI", shaking head to refuse something, or thumbs up to indicate approval. Actually, novel studies indicate that body gestures are helpful to understand hidden emotions or suppressed feeling. Additionally, combining facial expressions and body gestures in a bimodal manner have shown to be efficient for recognizing human nonverbal behavior [155], analyzing the correlations between gesture and emotions, and combining the categories in a multimodal manner. This research area lines up with affect recognition field which contains wider range of modalities including affect in written language, facial display, and physical activity. Affect computing (AC) [156] is the field that develops techniques able to recognize human emotions. In affective recognition, information is combined from facial expressions, human's behavioral, and affect states in order to predict the emotional feelings. In this regard, the significant contribution of psychologists and linguistics have impacted the progress in AC for the purpose of understanding human affect perception [157]. To this end, new challenges have been introduced including the necessity to study the correlations between the behavioral motions such as facial, head, and gestures as well as the importance to consider the temporal correlations between the different modalities.

The main contributions of this paper is to provide to the emotion analysis community two new emotion analysis models able to:

[1] https://appen.com/figure-eight-is-now-appen/

(1) Track emotions in children-directed texts ”fairy tales” using two novel topic models: Latent-based Smoothed Beta-Liouville and Smoothed Beta-Liouville Emotion Term Model that we have developed.

(2) Incorporate modeling unknown documents in a Bayesian folding-in way along with estimating the document-topic distribution.

(3) Detect affective states from facial and body recognition through a bimodal affect framework that combines facial expressions features, pose estimation, and hand gestures.

(4) Consider the correlations between different behavioral motions with latent topics using Latent SBL through exploiting the temporal phase detection into onset, apex, and offset.

The remainder of this paper is organized as follows. Section 2 reviews the related work mainly associated with emotion analysis from textual data and affect recognition from face and body. Section 3 presents a background introduction for the smoothed-based models and Bayesian folding-in method for PLSI. In Section 4, we present the novel Smoothed Beta-Liouville distribution and in Section 5 the Latent-based Smoothed Beta-Liouville. Following, in Section 6, we present the novel Smoothed Beta-Liouville kernels for PLSI and the emotion term model. Sections 7 and 8 demonstrate the experimental results for tracking emotions in fairy tales and affect recognition from face and body, respectively. Section 9 concludes this paper with conclusions remarks and future insights.

5.2 Related work

5.2.1 Emotion analysis from textual data

Emotion analysis from text has gained recently increased attention from linguistics, computer science, and psychologists. Text-based emotion recognition is one of the most important topics in NLP research area. For instance, emotions expressed in text format could be communicated through different modalities such as e-mails, letters, social media, and books. Methods of emotion detection rely mainly on bag of words modeling and statistical approaches that consider embedding words including word2vec [158], PAS-CBOW [159], and other word vector representations [160]. Traditional emotion classification models have been based on the common text classifiers like Support

Vector Machines (SVM), Naive Bayes, and Logistic regression. Afterwards, popular deep learning models, in particular, Convolutional Neural Network (CNN) [161], Recurrent Neural Network (RNN) [162], Long Short-Term Memory (LSTM) networks [163], and Bi-directional Long Short-Term Memory (Bi-LSTM) [164] offer a competitive performance for emotion analysis.

Despite the much focus on text-based emotion recognition, little is directed to fairy tales detection. The most related direction to our work is oriented to emotion analysis of books which includes fairy tales, fables, novels, romances, and epics. Emotion analysis of books has an outstanding importance as much as emotion recognition from social media. In fact, tracking emotions from books could be considered in several applications including social analysis (*i.e.* analyzing the distribution of words related to women, race, and religion), summarization, and analyzing persuasion tactics. In this regard, Mohammad [165] considers tracking emotions in mail and books where the author studied the emotional difference between man and woman in work-place mail and shows how fairy tales have more emotions densities than novels. Volkova et al. [166] explored the emotional perception of fairy tales through a new annotation scheme. Acerbi et al. [167] performed an emotion extraction model using three emotion detection tools (WNA, LIWC, HED) on Google Books n-gram corpus.

All the above works directed to emotion analysis of books just focus on studying the density concept of emotions [165, 166] and detecting emotion states [167] using classical detection tools. Little attention is paid to modeling text occurrence information for emotion recognition in fairy tales. To address these mentioned issues, our attention will be oriented to topic modeling because of its strength and flexibility. Particularly, Bao et al. [168] have extended Latent Dirichlet Allocation (LDA) to model the connection between words and emotions using an emotion-topic model where they add an additional emotion generation layer to LDA. However, LDA fails to generate accurate topics over short texts. To address the sparsity in feature space when modeling short text for emotion recognition, Pang et al. [169] proposed a weighted labeled topic model (WLTM). Wang et al. [170] introduced Topic-Over-Time (TOT) that captures time information with latent topics. Mei et al. [171] proposed Topic sentiment mixture model for sentiment analysis.

5.2.2 Affect Recognition from face and body

Affect expressions occur through combination of different non-verbal communications forms such as facial expressions, body posture, gestures, and eye gaze. Despite the large range of modalities, most of the research works [172] have focused on recognizing affective states from facial muscle actions (facial action units) rather than emotions from facial displays. Several studies have demonstrated the importance of integrating different modalities for human affect perception [173]. Only recently, some studies have considered bodily expressions and have shown their importance as much as facial expressions in detecting emotions [174]. Body language, for instance, gestures and posture, can conduct an important message about how humans communicate their feelings and emotions. Progressive attention is being made toward affect expressions through body language systems. In fact, potential applications of affect recognition entails commercial call services, intelligent automobile systems, video-gaming, and video surveillance systems. In addition, affect-related research work fields including psychiatric disorder, behavioral science, and neuroscience have been able to help in assisting patients for mental disease treatment. Actually, few databases have been devoted to analyze and recognize affective face and body emotional expressions. In this regard, Gunes and Piccardi [175, 155] introduced the FABO database, where they proposed a recognition system that uses feature vectors combining the upper body and facial expression and a set of standard classifiers (SVM, Adaboost, C4.5, HMM). Baltrusaitis et al. [176] proposed GEMEP-FERA database which is a subset of GEMEP corpus and they considered a hierarchical recognition system based on HMM. Further, Castellano et al. collected a body language database HUMAINE [177]. In their recognition model, they compared different models including 1-nearest-neighbor with dynamic time warping (DTW-1NN), decision tree, and Hidden Naive Bayes. Baveye et al. [178] created LIRIS-ACCEDE database which contains upper body videos annotated with six basic emotions.

The methods proposed in this paper differs significantly from the aforementioned approaches for different reasons. The proposed emotional body gesture recognition models were based on standard classifiers. However, the features extracted from different modalities are more complex and represent several challenges that couldn't be handled using basic AI algorithms. These challenges are, for instance, the correlation between features from different modalities and also the temporal properties of face and body gestures. Furthermore, our models are based on the nature of the data considering the textual and the visual features. For instance, we consider the extension of the smoothed Dirichlet distribution which is proposed for building topic models for text. Our purpose is to improve the Dirichlet distribution and to tackle the problems of data covariance structure with an alternative distribution namely the Beta-Liouville. In addition, despite the success of the previous works for predicting emotions, existing approaches consider only modeling the topic mixtures of known documents and ignored the unknown ones. In contrast, in this paper, we aim to develop a topic mixture model so-called Latent-based Smoothed Beta-Liouville (Latent SBL) and a Bayesian folding-in with Smoothed Beta-Liouville kernels for PLSI which incorporates modeling the unknown documents in a Bayesian folding-in way.

5.3 Background

This section describes the background of the models proposed and the motivations behind this choice. First, we start by introducing Smoothed Dirichlet and Smoothed Generalized Dirichlet. Second, we present PLSI and Bayesian folding-in.

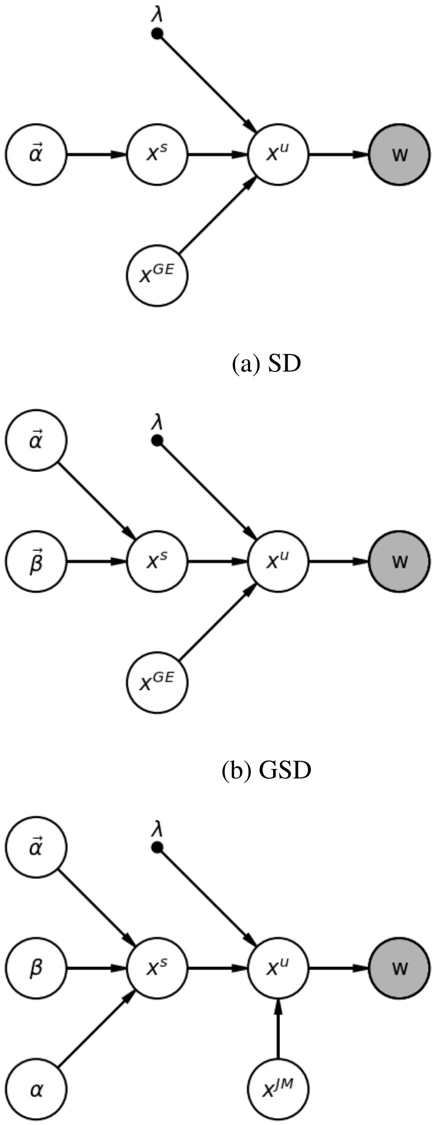
5.3.1 Smoothed-based models

In the context of building generative topic models for text, Nalapati et al. [82] proposed the smoothed Dirichlet (SD) distribution. SD distribution, a novel variant of the Dirichlet, was introduced as an alternative to the Dirichlet compound multinomial model [22] for text classification. For generating text, the probability density function of SD is given as follows:

$$p(\vec{X}^s|\vec{\alpha}) = \frac{\prod_{d=1}^{D}(x_d^s)^{\alpha_d - 1}}{\lambda \int_{\Delta^s} \prod_{d=1}^{D}\{\lambda x_d^u + (1-\lambda)x_d^{GE}\}^{\alpha_d - 1} dX^u}, \tag{105}$$

where $\vec{X}^s$ is the smoothed proportions vector defined by:

$$\vec{X}^s = \lambda \vec{X}^u + (1-\lambda)\vec{X}^{GE}, \tag{106}$$



(a) SD

(b) GSD

(c) SBL

Figure 5.1: Graphical representation of SD, GSD, and SBL distributions

where $\vec{X}^u$ is a proportions vector defined by $\vec{X}^u = (w_1/\sum_d w_d, \ldots, w_D/\sum_d w_d)$ (w_d is a the frequency of word in a document), $\vec{X}^{GE}$ is the general proportions of words estimated from the entire document, and $0 < \lambda < 1$ is a smoothing parameter. The generation of smoothed proportions for the smoothed Dirichlet distribution is displayed in Figure 5.1 (a).

The smoothed Dirichlet distribution is proposed under a subset of the entire simplex Δ where the smoothed proportions of words are defined:

$$\Delta^s = \{\vec{X}^s\} = \{\lambda \vec{X}^u + (1-\lambda)\vec{X}^{GE} | \vec{X}^u \in \Delta\}, \tag{107}$$

Integrating the normalizer of SD distribution with regards to the new compressed domain was simplified using Stirling's approximation of Gamma function which gives:

$$p(\vec{X}^s|\vec{\alpha}) = \frac{\sum_{d=1}^{D} \alpha_d^{\sum_{d=1}^{D} \alpha_d}}{\prod_{d=1}^{D} \alpha_d^{\alpha_d}} \prod_{d=1}^{D} (x_d^s)^{\alpha_d - 1}, \tag{108}$$

where $\vec{\alpha}$ is the SD parameter.

SD was successfully applied also for information retrieval [89], word embedding [13], and image categorization [14] where it outperforms SVM, DCM, CNN, and LDA. However, the smoothed Dirichlet suffers from the same disadvantages of Dirichlet distribution which relies on the facts that features with the same mean must have the same covariance and it has a restrictive negative covariance structure. In this regard, Najar et al. [16] proposed a smoothed variant of the generalized Dirichlet distribution and a generalization of the smoothed Dirichlet. The Smoothed Generalized Dirichlet (SGD) was proposed for emotion detection including disaster tweets related emotions and pain intensity estimation. SGD is based on the concept of neutrality that offers a large structure for the covariance and has one extra free parameter which makes the distribution more flexible than SD. SGD was proposed for modeling count data where the probability of generating a smoothed count vector is defined by:

$$
\begin{aligned}
p(\vec{X}^s|\vec{\alpha},\vec{\beta}) &= \prod_{d=1}^{D-1} \frac{(\alpha_d+\beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}} \\
& \quad (x_d^s)^{\alpha_d-1}\big(1-\sum_{k=1}^{d} x_k^s\big)^{\gamma_d}
\end{aligned}
\tag{109}
$$

where $\vec{\alpha}=(\alpha_1,\ldots,\alpha_{D-1})$ and $\vec{\beta}=(\beta_1,\ldots,\beta_{D-1})$ are the parameters of Smoothed Generalized Dirichlet that characterize the shape of the distribution, and $\gamma_d=\beta_d-\alpha_{d+1}-\beta_{d+1},\ d=1,\ldots,D-2$ and $\gamma_{D-1}=\beta_{D-1}-1$.

Under the above density function, we demonstrate in Figure 5.1 (b) how smoothed words are generated with SGD parameters $\vec{\alpha}$ and $\vec{\beta}$.

5.3.2 Bayesian folding-in for PLSI

Probabilistic Latent Semantic Indexing (PLSI) [179] was initially proposed for document indexing and information retrieval. PLSI models documents as a mixture of latent topics where its parameters including the topic-word associations and document-topic mixtures are learned using Expectation-Maximization (EM) algorithm. The PLSI model is a latent variable model that represents the co-occurrence documents d and words w as a mixture of z latent topics:

$$
p(d,w)=p(d)\sum_{z}p(w|z)p(z|d) \tag{110}
$$

where $p(w|z)$ are the word distributions and $p(z|d)$ are the weights of distributions.

However, PLSI is a non-generative model and can't be used for computing the representation of a new unknown document. For that, a Folding-in approach is proposed for obtaining the topic mixture proportions $p(w|z)$. The folding-in procedure estimates as well the model's parameters using EM algorithm where $p(z|d,w)$ is learnt in the E-step while $p(w|z)$ and $p(d|z)$ are estimated during M-step. Though, the problem with folding-in of new document is the fact that topic mixtures of known documents are ignored. Hence, this model leads to an insignificant representation of the novel extended collection of documents. To deal with this issue, Bayesian folding-in [180] was proposed to model PLSI parameters in a Bayesian way for information retrieval. For Bayesian folding-in, instead of maximizing the likelihood function of the word vector, the posterior is maximized. To cope with modeling the topic mixture of the known documents in collection, the prior is defined as a kernel density estimate using Dirichlet distribution “Dirichlet kernels”.

5.4 Smoothed Beta-Liouville distribution

In this chapter, we propose a smoothed Beta-Liouville distribution: a novel probabilistic approach for modeling textual data. First, we introduce the Liouville family and the Beta-Liouville distribution. Then, we present the steps that leads to define the new smoothed distribution.

The Liouville family of distributions is defined with positive parameters $(\alpha_1, \dots, \alpha_d, \zeta)$ as follows:

$$\begin{aligned} p(\vec{X}|\alpha_1, \dots, \alpha_D, \zeta) &= f(u|\zeta)\frac{\Gamma(\sum_{d=1}^{D}\alpha_d)}{u^{\sum_{d=1}^{D}\alpha_d - 1}} \\ & \prod_{d=1}^{D}\frac{x_d^{\alpha_d-1}}{\Gamma(\alpha_d)}, \end{aligned} \tag{111}$$

where $u = \sum_{d=1}^{D} x_d$ and $f(u|\zeta)$ is the density generator of u with parameters ζ.

The probability density function of the Liouville family is defined in the simplex
$\Delta = \{(x_1, \dots, x_D); \sum_{d=1}^{D} x_d \leq a\}$ if and only if the density generator $f(.)$ is defined in $[0, a]$. For this purpose, we choose the Beta distribution defined in $[0, 1]$ in view of modeling proportional data. The probability density function of Beta distribution is given by:

$$f(u|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}u^{\alpha-1}(1-u)^{\beta-1}, \tag{112}$$

In dimension D, a categorical data vector $\vec{X}$ is defined in the simplex $\Delta = \{(x_1, \dots, x_D), \sum_{d=1}^{D} x_d \leq 1\}$. We present a transformation on the simplex that contains smoothing representation for the categorical data which can be defined as the following expression:

$$\Delta^s = \{(x_1, \dots, x_D), \lambda x_d^u + (1-\lambda)x_d^{JM} | x_d \in \Delta\}, \tag{113}$$

where λ is a smoothing parameter and x^{JM} is a normalization vector estimated from the entire dataset using Jelinek-Mercer (JM) smoothing [81]. Defining the smoothed proportions vector $\vec{X}^s \in \Delta^s$ from the Beta-Liouville distribution gives:

$$p(\vec{X}^s|\Theta) = \frac{1}{Z^s}(u^s)^{\alpha-\sum_{d=1}^{D}\alpha_d}(1-u^s)^{\beta-1}\prod_{d=1}^{D}(x_d^s)^{\alpha_d-1}, \tag{114}$$

where $\Theta = (\alpha_1, \ldots, \alpha_D, \alpha, \beta), u^s = \sum_{d=1}^{D} x_d^s$, and Z^s is the normalizer that guarantees that the probabilities add up to 1 defined as follows:

$$Z^s = \int_{\Delta^s}(u^s)^{\alpha-\sum_{d=1}^{D}\alpha_d}(1-u^s)^{\beta-1}\prod_{d=1}^{D}(x_d^s)^{\alpha_d-1}d\vec{X}^s, \tag{115}$$

Considering the fact that the smoothed simplex is a subset of the whole simplex Δ, it is clearly incorrect to address the Beta-Liouville normalizer given by:

$$\begin{aligned} Z &= \int_{\Delta}(u)^{\alpha-\sum_{d=1}^{D}\alpha_d}(1-u)^{\beta-1}\prod_{d=1}^{D}(x_d)^{\alpha_d-1}d\vec{X} \\ &= \frac{\prod_{d=1}^{D}\Gamma(\alpha_d)\Gamma(\beta)\Gamma(\alpha)}{\Gamma(\sum_{d=1}^{D}\alpha_d)\Gamma(\alpha+\beta)}, \end{aligned} \tag{116}$$

Thus, we need to define an analytical form for the smoothed normalizer defined in the compressed domain Δ^s:

$$\begin{aligned} Z^s = &\int_{\Delta^s}(u^s)^{\alpha-\sum_{d=1}^{D}\alpha_d} \\ &(1-u^s)^{\beta-1}\prod_{d=1}^{D}(x_d^s)^{\alpha_d-1}d\vec{X}^s, \end{aligned} \tag{117}$$

where $u^s = \sum_{d=1}^{D} x_d^s = \lambda u + (1-\lambda)u^{JM}$ and $u^{JM} = \sum_{d=1}^{D} x_d^{JM}$. Thus, we have:

$$\begin{aligned} Z^s \quad &= \int_{\Delta} (\lambda u + (1-\lambda)u^{JM})^{\alpha - \sum_{d=1}^{D} \alpha_d} (1 - \lambda u + (1-\lambda)u^{JM})^{\beta - 1} \\ & \prod_{d=1}^{D} (\lambda x_d + (1-\lambda)x_d^{JM})^{\alpha_d - 1} \lambda d\vec{X}, \end{aligned} \tag{118}$$

Taking cue from the smoothed Dirichlet normalizer defined in [82], we define the approximation of the Beta-Liouville normalizer Z^s as:

$$Z^s \quad = \frac{\prod_{d=1}^{D} \Gamma_a(\alpha_d)\Gamma_a(\alpha)\Gamma_a(\beta)}{\Gamma_a(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)}, \tag{119}$$

where $\Gamma_a(.)$ is an approximation to $\Gamma(.)$ under the Stirling's approximation defined as follows:

$$\Gamma(\alpha) \approx e^{-\alpha} \alpha^{\alpha - 1/2} \sqrt{2\pi} (1 + \frac{1}{12\alpha} + O(\frac{1}{\alpha^2})), \tag{120}$$

Taking into consideration the fact that we need a bounded normalization solution for the smoothed distribution, we follow the simplification considered in [82] to redefine the Gamma approximation function as:

$$\Gamma_a(\alpha) \simeq e^{-\alpha} \alpha^{\alpha}, \tag{121}$$

Note that we choose simply to ignore the unbounded terms in Stirling's approximation that yields to a mathematical simpler solution and a closed form solution to maximum likelihood estimation. Combining equations 119 and 121 gives:

$$\begin{aligned} Z^s \quad &\simeq \quad \frac{\prod_{d=1}^{D} e^{-\alpha_d} \alpha_d^{\alpha_d} e^{\alpha} \alpha^{\alpha} e^{-\beta} \beta^{\beta}}{e^{-\sum_d \alpha_d} (\sum_d \alpha_d)^{\sum_d \alpha} e^{-(\alpha+\beta)} (\alpha+\beta)^{\alpha+\beta}} \\ &\simeq \quad \frac{\prod_{d=1}^{D} \alpha_d^{\alpha_d} \alpha^{\alpha} \beta^{\beta}}{S^S (\alpha+\beta)^{\alpha+\beta}}, \end{aligned} \tag{122}$$

where $S = \sum_{d=1}^{D} \alpha_d$. Note that defining the new distribution depends on the Beta-Liouville normalizer, we eventually present the novel Smoothed Beta-Liouville (SBL) distribution as follows:

$$p(\vec{X}^s|\Theta) = \frac{(\alpha+\beta)^{\alpha+\beta} S^S}{\alpha^\alpha \beta^\beta \prod_{d=1}^D \alpha_d^{\alpha_d}} (u^s)^{\alpha-S} (1-u^s)^{\beta-1} \prod_{d=1}^D (x_d^s)^{\alpha_d - 1}, \tag{123}$$

With $\alpha = \sum_{d=1}^D \alpha_d$ and $\beta = \alpha_{D+1}$, the Smoothed Beta-Liouville distribution is reduced to the Smoothed Dirichlet distribution with parameters $(\alpha_1, \ldots, \alpha_{D+1})$. We compare in Figure 5.1 the generation of the smoothed word proportions using the proposed SBL distribution with SD and GSD. We present the graphical representations for the three models where SBL has more characterizing parameters namely $\vec{\alpha}$, β, and α.

5.5 Latent-based Smoothed Beta-Liouville

We propose, in this section, a mixture modeling based approach. A generative topic model constituted by a mixture of SBL distributions, so-called Latent-based Smoothed Beta-Liouville (Latent SBL). In what follows, we consider a mixture of K Smoothed Beta-Liouville distributions. Given a collections of N documents $\mathcal{D} = \{d_1, \ldots, d_N\}$ where each document is represented by the co-occurrence of words x_d in a vocabulary of size D, the Latent SBL represents documents as a mixture of K SBL distributions which depict the latent topics. For this approach, the log-likelihood function can be written as follows:

$$\begin{aligned} l(\Theta) &= \sum_{i=1}^N \sum_{j=1}^K \hat{w}_{ij} \Big[(\alpha_j + \beta_j) \log(\alpha_j + \beta_j) + S_j \log(S_j) \\ &+ (\alpha_j - S_j) \log(\sum_{d=1}^D x_{id}^s) + (\beta_j - 1) \log(1 - \sum_{d=1}^D x_{id}^s) \\ &- \alpha_j \log(\alpha_j) - \beta_j \log(\beta_j) - \sum_{d=1}^D \alpha_{dj} \log(\alpha_{dj}) \\ &+ \sum_{d=1}^D (\alpha_{dj} - 1) \log(x_{id}^s) + \log(p_j) \Big], \end{aligned} \tag{124}$$

where p_j represents the mixing proportions which are positive and sum to 1 and $\hat{w}_{ij}$ is the posterior probability given by:

$$\hat{w}_{ij} = \frac{p_j p(\vec{X}^s|\Theta)}{\sum_{j=1}^{K}(p_j p(\vec{X}^s|\Theta))}, \tag{125}$$

Seeking to find the maximum likelihood estimates (MLE) of Θ that optimize the above log-likelihood function, we employ the Newton-Raphson algorithm which is defined as the following:

$$\Theta_{new} = \Theta_{old} - H^{-1}g, \tag{126}$$

where g is the gradient vector and H^{-1} is the inverse of Hessian matrix which request the calculation of second-order derivatives. The gradient is therefore defined as:
for $d = 1, \dots, D, j = 1, \dots, K$,

$$g_{\alpha_{jd}} = \sum_{i=1}^{N} \hat{w}_{ij} \Big[\log(x_{id}^s) + \log(\sum_{d=1}^{D} \alpha_{jd}) - \log(\alpha_{jd}) \Big], \tag{127}$$

$$g_{\alpha_j} = \sum_{i=1}^{N} \hat{w}_{ij} \Big[\log(\alpha_j + \beta_j) + \log(\sum_{d=1}^{D} x_{id}^s) - \log(\alpha_j)], \tag{128}$$

$$g_{\beta_j} = \sum_{i=1}^{N} \hat{w}_{ij} \Big[\log(\alpha_j + \beta_j) + \log(1 - \sum_{d=1}^{D} x_{id}^s) - \log(\beta_j)], \tag{129}$$

The Hessian matrix is block diagonal and its inverse is represented as follows:

$$H^{-1} = \text{block-diag}\{H(l(\alpha_{j1}, \dots, \alpha_{jD}))^{-1}, H(l(\alpha_j, \beta_j))^{-1}\}, \tag{130}$$

where

$$H(l(\alpha_{j1}, \dots, \alpha_{jD})) = \sum_{i=1}^{N} \hat{w}_{ij} \begin{bmatrix} C_{\alpha_d} - \frac{1}{\alpha_{j1}} & \dots & C_{\alpha_d} \\ \vdots & \ddots & \vdots \\ C_{\alpha_d} & \dots & C_{\alpha_d} - \frac{1}{\alpha_{jD}} \end{bmatrix}, \tag{131}$$

and

$$H(l(\alpha_j, \beta_j)) \;=\; \sum_{i=1}^{N} \hat{w}_{ij} \begin{bmatrix} C_{\alpha\beta} - \frac{1}{\alpha_j} & C_{\alpha\beta} \\ C_{\alpha\beta} & C_{\alpha\beta} - \frac{1}{\beta_j} \end{bmatrix}, \tag{132}$$

where $C_{\alpha_d} = \frac{1}{\sum_{d=1}^{D} \alpha_{jd}}$ and $C_{\alpha\beta} = \frac{1}{\alpha_j + \beta_j}$ are constant terms. Then, we use for the inverse of the matrix $H(l(\alpha_{j1}, \ldots, \alpha_{jD})$ the following formula:

$$H(l(\alpha_{j1}, \ldots, \alpha_{jD})^{-1} = \text{diag}[h]^{-1} + \delta^* a^* a^{*tr}, \tag{133}$$

where the $(.)$ function places a given vector on the diagonal of a matrix, h is a column vector containing the non-constant terms from the diagonal of the Hessian matrix, a^* is a column vector of ones, and $\delta^* = -\sum_{i=1}^{N} \hat{w}_{ij} C_{\alpha_d} (1 + \sum_{i=1}^{N} \hat{w}_{ij} C_{\alpha_d} \sum_{d=1}^{D} \frac{1}{h_d})^{-1}$.

We develop the complete clustering algorithm using an Expectation-Maximization approach where we evaluate the posterior probability defined in equation 125 and next we maximize the parameters of Latent SBL using maximum likelihood estimator.

5.6 Smoothed Beta-Liouville Kernels for PLSI

A probabilistic Latent Semantic Indexing (PLSI) is a basic aspect model proposed for latent topic models [179] which represents a document as a mixture of latent topics. Several research works have extended the PLSI model to overcome its drawbacks including the overfitting problem. The Latent Dirichlet Allocation (LDA) was proposed [181] to model documents as mixture of topics drawn from a Dirichlet distribution which is considered as a generalization of PLSI approach. Other alternatives which use log-normal prior distributions include correlated topic models [182] and dynamic topic models [183]. Further, undirected graphical models [184] have improved PLSI approach using contrastive divergence such as undirected PLSI and Rate Adapting Poisson model. A direct improvement approach for the PLSI is the use of Fisher kernels to learn document similarities [185] and in another work a kernel density estimate was used as prior in Bayesian folding-in procedure [180] which also captures the dependencies between topics. This approach uses the maximum a posteriori instead of the maximum likelihood estimation method considered in the original PLSI approach. However, the used kernel was a Dirichlet density which makes the model susceptible also to overfitting problem. For that purpose, we propose in this work to extend the Bayesian folding-in approach and to use as a prior the Smoothed Beta-Liouville kernels.

Using the same notation defined in the previous section, PLSI folding-in estimates the mixture proportions $\vec{\theta}_q$ of topics $z \in Z$ for a new document d_q represented by new word vector $\vec{X}_q = (x_{q1}, \dots, x_{qD})$ where the parameters of the model are the document-topic distributions $\vec{\theta} = [\theta_{ij} = P(z_j|d_i)]_{i=1,\dots,N, j=1,\dots,K}$ and the topic-word associations $\vec{w} = [w_{dj} = P(x_d|z_j)]_{d=1,\dots,D,j=1,\dots,K}$. The maximum a postoriori (MAP) requires maximizing the posterior distribution $P(\vec{\theta}_q|\vec{X}_q, \vec{\theta}, \vec{X})$ to estimate the topic mixtures by running the Expectation-Maximization algorithm. We define the posterior distribution for a new document d_q as follows:

$$P(\vec{\theta}_q|\vec{X}_q, \vec{\theta}, \vec{X}) \propto P(\vec{X}_q|\vec{\theta}_q, \vec{\theta}, \vec{X}) P(\vec{\theta}_q|\vec{\theta}, \vec{X}), \tag{134}$$

where $P(\vec{X}_q|\vec{\theta}_q, \vec{\theta}, \vec{X})$ is the word likelihood and $P(\vec{\theta}_q|\vec{\theta}, \vec{X})$ is the topic prior. Assuming here that the words are independent, so the likelihood can be defined as:

$$P(\vec{X}_q|\vec{\theta}_q, \vec{\theta}, \vec{X}) = \prod_{d=1}^{D} \sum_{j=1}^{K} P(x_d|z_j) P(z_j|q) = \prod_{d=1}^{D} \sum_{j=1}^{K} w_{dj} \theta_{qj}, \tag{135}$$

We define the topic prior as a kernel density estimate based on Smoothed Beta-Liouville kernels where the topics have non-negative components and sum to one which makes the SBL distribution a suitable density function over the smoothed simplex. Defining the SBL kernel, we introduce a smoothing parameter η for each document-topic mixture vector. Thus, the kernel density prior is defined as follows:

$$P(\vec{\theta}_q|\vec{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \text{SBL}(\vec{\theta}_q^s|\alpha_s, \beta_s, \vec{\alpha}(\vec{\theta}_i)), \tag{136}$$

where $\vec{\theta}_q^s$ is the smoothed document-topic defined in the smoothed simplex Δ^s, $\vec{\alpha}(\vec{\theta}_i) = \frac{1}{\eta}\vec{\theta}_i + 1$,

$\alpha_s = \frac{1}{\eta}\alpha + 1$, and $\beta_s = \frac{1}{\eta}\beta + 1$ are the smoothing functions of the parameters of SBL.

To simplify the maximization of the posterior distribution with respect to $\vec{\theta}^s_q$, we define new hidden variables such as a binary variable $\vec{y}_d \in \{0,1\}^K$ which indicates if a topic z_j contains a word x_d and an indicator variable $\vec{r} \in \{0,1\}^N$ that indicates which SBL distribution represents a document-topic mixture of the new document. Thus, we define the modified log-posterior function of the new model as:

$$\begin{aligned} \log[P(\vec{\theta}_q|\vec{X}_q, \vec{y}, \vec{r}, \vec{X})] &= \log[P(\vec{X}_q, \vec{y}|\vec{\theta}_q, \vec{X})P(\vec{\theta}_q, \vec{r}|\vec{\theta}, \vec{X})] \\ &= \left[\prod_{d=1}^{M}\sum_{j=1}^{K} y_{dj}[\log w_{dj} + \log\theta_{qj}]\right] \\ &+ \sum_{i=1}^{N} r_i[\log\frac{1}{N} + \log \mathrm{SBL}(\vec{\theta}^s_q|\alpha_s, \beta_s, \vec{\alpha}(\vec{\theta}_i))], \end{aligned} \tag{137}$$

The parameters of PLSI are estimated through an EM algorithm where the posteriors for the hidden variables are defined in the E-step and the document-topic mixtures for the new document are updated in the M-step.

E-Step: we define the following posteriors of the hidden variables.

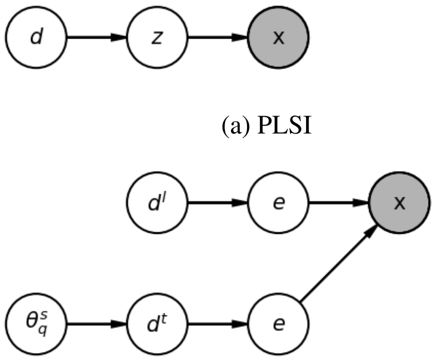
$$P(y_{dj} = 1|x_d, \vec{\theta}^{(t)}_q, \vec{X}) = \frac{w_{dj}\theta^{(t)}_{qj}}{\sum_{j'=1}^{K} w_{dj'}\theta_{qj'}} = g_{dj}, \tag{138}$$

$$P(r_i = 1|\vec{\theta}^{s(t)}_q, \vec{\theta}) = \frac{\mathrm{SBL}(\vec{\theta}^{s(t)}_q|\alpha_s, \beta_s, \vec{\alpha}(\vec{\theta}_i))}{\sum_{i'=1}^{N}\mathrm{SBL}(\vec{\theta}^{s(t)}_q|\alpha_s, \beta_s, \vec{\alpha}(\vec{\theta}'_i))} = h_i, \tag{139}$$

M-step: we substitute the posteriors defined in E-step into Equation 138 then we maximize the log-posterior w.r.t. $\vec{\theta}_q$ under the condition of $\sum_{j=1}^{K}\theta_{qj} = 1$ which gives the following formula:

$$\theta^{s(t+1)}_{qj} = \frac{\sum_{d=1}^{D} g_{dj} + \frac{1}{\eta}\sum_{i=1}^{N} h_i\theta^{s(t)}_{ij}}{D + \frac{1}{\eta}}, \tag{140}$$

In this work, we adapt the proposed Bayesian folding-in approach to text-based emotion recognition. We consider a training data collection containing N^l documents denoted by $\mathcal{D}^l = \{(d^l_1, e^l_1),$ $\ldots, (d^l_{N^l}, e^l_{N^l})\}$ where d^l_i and e^l_i is the i-th training text document and its corresponding emotion label respectively. The unlabeled target documents set is denoted by $\mathcal{D}^t = \{d^t_1, \ldots, d^t_Q\}$ where d^t_i



(a) PLSI

(b) SBL-ETM

Figure 5.2: Graphical model for PLSI and SBL-ETM

is the i-th unlabeled document in the test corpus containing Q documents. The objective here is to assign an emotion label to each new document in the testing corpus. First, we assume that the number of latent topics is equal to emotion categories in the whole corpus. Second, we consider that emotion-word $p(w|e)$ probabilities are the same for both collections and the probability of generating training documents with a specified emotion category $p(d^l|e)$ is different from the probability of generating a new document form the testing collection with same emotion category $p(d^t|e)$. Finally, after training the model learning, we predict the new document label based on the posterior distribution $p(e|d^t)$ as follows:

$$\begin{aligned} e &= \text{argmax}\, p(e|d^t) \\ &= \text{argmax}\, \vec{\theta}_q^s \end{aligned} \tag{141}$$

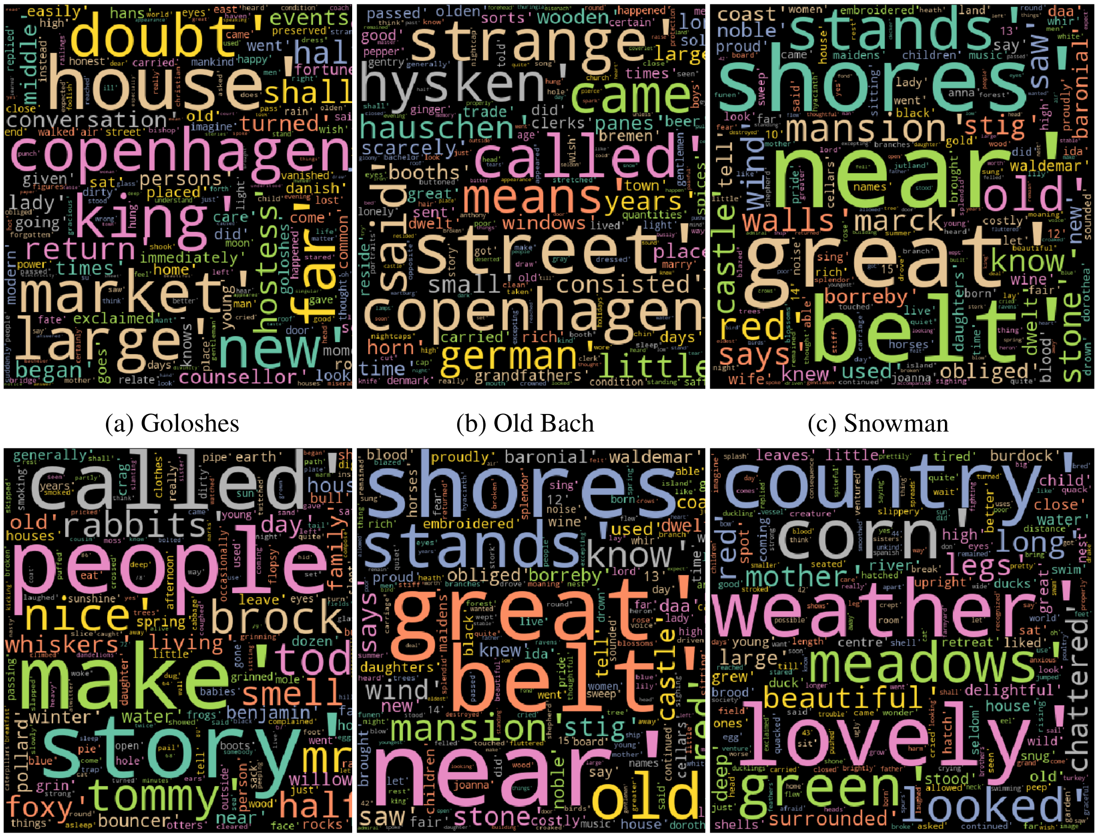
The new emotion-term model is shown in Figure 5.2 where the training and testing domains are incorporated in the so-called Smoothed Beta Liouville Emotion Term model (SBL-ETM).

5.7 Tracking emotions in fairy tales

In this section, we introduce the first application where we track emotions in fairy tales using our proposed approaches. We consider the Fairy tales dataset [186] which includes a collection of children's stories written by Potter, the Brothers Grimm, and Andersen. The corpus has been independently marked by two annotators with a *primary emotion* and a *mood* label for each sentence which includes the feeling of an audience listening to a story being read. The primary and mood annotations were combined into a set of 8 affect categories for each sentence including: neutral, angry, disgusted, fearful, happy, sad, positively surprised, and negatively surprised. Each collection contains a set of stories: Grimms including 80 stories with a total of 5352 sentences, Potter (18 stories, 1926 sentences), and Andersen (77 stories, 7984 sentences). Each line in the text dataset consists of Sentence ID, Emotion label for annotator A and B, Mood label for annotator A and B, and the sentence. In this work, we evaluate our proposed models on three different experiments: tracking emotions on short stories, on the three collections, and on new stories from collections. For the three experiments, we consider the emotion label marked by annotator A for tracking emotions purposes.

First, for preprocessing the text data, we eliminate the stop words as well as the infrequent words to create a vocabulary with different feature sizes. Next, we represent each sentence by a vector of co-occurrences of words which are used for clustering with the proposed model. We show in Figure 5.3, examples of wordcloud generated from six different fairy tales.

For experiment 1, we evaluate the Latent SBL on six different stories where Goloshes, The tale of Mr Tod, Old Bach, Ugly duc, The wind, and Snowman consist of 495, 314, 212, 164, 114, and 114 sentences respectively. We show in Table 5.1 the evaluation results of Latent SBL on these short stories and comparing with the results of SGD and SD. The tracking scores show the robustness of Latent SBL for clustering emotion categories in short stories. In fact, this type of data suffers from short text issues such as the sparseness problem and the unmeaningful content which is often complicated to predict the underlying emotion information. Also, small datasets has the challenge of unsufficient data for learning a model. From Table 5.1, we demonstrate the out-performance of the novel Latent SBL with regards to SGD and SD where the properties of Beta-liouville are proved to make it the best adequate distribution for clustering such type of data and for emotion categorization as well. We mention also the significance of incorporating the latent topics in our model which identify the topical structure of the textual data and shows higher quality in the case of limited size corpora.



(a) Goloshes

(b) Old Bach

(c) Snowman

(d) The tale of Mr Tod

(e) The wind

(f) Ugly duc

Figure 5.3: Word-cloud generated from six different Fairy tales: Goloshes, Old Bach, Snowman, The tale of Mr Tod, The wind, and Ugly duc.

Table 5.1: Comparing tracking scores of Latent SBL for different fairy tales (sentence-level) with related smoothed-based models

Story	Model	Tracking scores
Goloshes	Latent SBL	83.63
	SGD	79.39
	SD	54.95
The tale of Mr Tod	Latent SBL	78.91
	SGD	67.41
	SD	60.10
Old Bach	Latent SBL	79.71
	SGD	68.39
	SD	53.11
Ugly duc	Latent SBL	72.60
	SGD	63.69
	SD	57.66
The wind	Latent SBL	83.03
	SGD	73.52
	SD	69.29
Snowman	Latent SBL	80.70
	SGD	78.40
	SD	69.29

In Table 5.2, we evaluate Latent SBL and SBL-ETM through two different experimental setups. The Latent SBL is applied on three collections namely Potter, Grimms, and Andersen over a clustering framework while for SBL-ETM, we train the model on documents from different collections containing 30 stories (Briar rose, Dog and sparrow, Fisherman and his wife, Golden bird, etc.) from which we construct a vocabulary of 5302 words. Then, Potter, Grimms, and Andersen collections are given as the testing sets where we predict the emotion categories on sentence level. The performance of both Latent SBL and SBL-ETM are higher than the related smoothed-based models (SGD, SD) for all collections. From Table 5.2, we notice that the two proposed latent topic models perform similarly. This can be explained by the fact that they are both based on the Beta-Liouville distribution. Additionally, we show the strength of our proposed approaches when dealing with large corpora (1926, 5352, 7984 sentences for Potter, Grimms, and Andersen, respectively), with highly dimensional data vectors (5302 co-occurrence features), and when predicting new unknown documents using SBL kernels for PLSI in a Bayesian folding-in way.

Table 5.2: Comparing tracking scores of Latent SBL for different fairy tales collections with related smoothed-based models

Collection	Model	Tracking scores
Potter (18 stories)	SBL-ETM	73.91
	Latent SBL	74.23
	SGD	50.20
	SD	47.81
Grimms (80 stories)	SBL-ETM	52.52
	Latent SBL	55.22
	SGD	52.26
	SD	50.06
Andersen (77 stories)	SBL-ETM	72.49
	Latent SBL	73.67
	SGD	65.33
	SD	58.31

5.8 Affect recognition from Face and Body

In this section, we address the recognition of human nonverbal emotional states. We consider for this study a bimodal face and body gesture database (FABO) [155] proposed for affective behavior analysis. The FABO database consists of recordings of 23 subjects that simultaneously performed face and body gestures with two cameras: the face camera and the body camera. Face and body videos of the FABO database have been labeled by six annotators into 12 emotional states including non-basic affective expression such as anxiety, boredom, uncertainty, puzzlement, and neutral/negative/positive surprise, in addition to the basic emotions of fear, anger, disgust, sadness, and happiness. Figure 5.4 shows sequence samples from FABO dataset from both body and face cameras which includes non-basic facial expressions and their corresponding body gestures. For this purpose, we adapt our model to cope with the mutlimodal data. We present in the first subsection the data preprocessing technique. Following, we display the different feature quantization methods considered for face and body. Then, we provide the affect recognition approach proposed in this work.

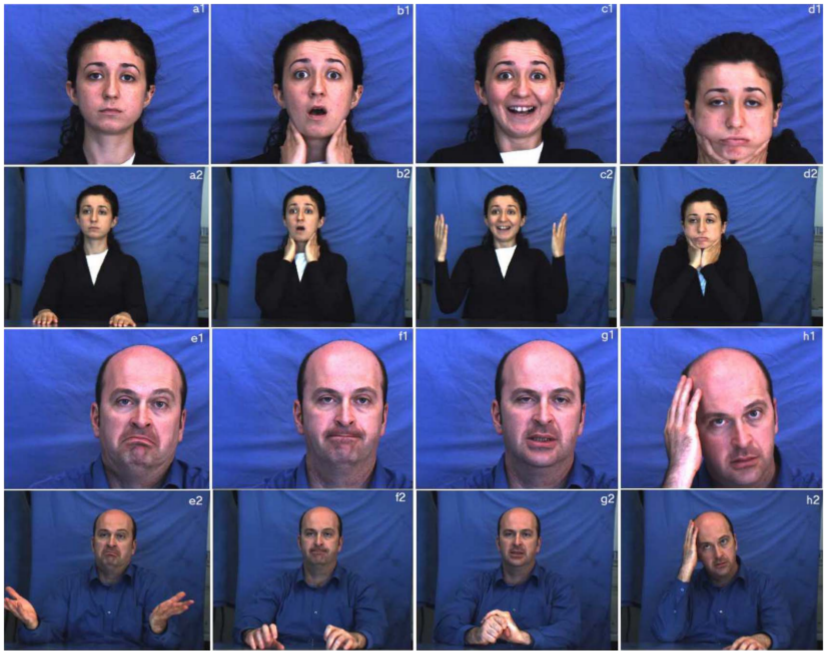


Figure 5.4: Sequences samples from FABO dataset obtained from body and face cameras. (a1–h1) non basic facial expressions and (a2–h2) their corresponding body gestures. (a) Neutral. (b) Negative surprise. (c) Positive surprise. (d) Boredom. (e) Uncertainty. (f–g) Anxiety. (h) Puzzlement.

5.8.1 Data pre-processing

In this work, we considered affect recognition on frame level. Therefore, we started by video preprocessing and frames extraction. In the work of Gunes et al. [175], for recognizing affect states, a temporal structure of facial movement is approximated by a sequence of four temporal phases called neutral, onset, apex, and offset. They showed the importance of temporal dynamics for interpreting emotional expressions. Neutral phase occurs when there are no sign of muscular action. The *onset* level is the phase of appearing of face changes while the *apex* is when the intensity of movement reaches a stable level followed by the relaxation of the muscular action that takes place in the *offset* phase. The body gesture consists of similar temporal factors which involves five phases: preparation, prestroke, stroke, poststroke, retraction. In this work, we consider the onset-apex-offset time markers for both facial and body gestures. Following experiments of [175], all the frames in the emotional sequence are classified into the temporal segment *apex* that is considered to be between frames number 37 and 57. For this reason, we extracted, first, all the frames from the video sequences and then we selected only the apex frames from which we derived the features.

5.8.2 Feature quantization

In our system, we used a combination of facial expression and body features.

Facial features For face model, we considered only the video obtained from face camera. In this work, we used a face recognition [2] library based on the dlib toolbox which is an open source machine learning software written in C++ [3]. Initially, we detected the position and boundaries of faces from the apex frame using face detection algorithm. Then, we extract 128-dimensional feature vectors using CNN algorithm. Figure 5.5 displays examples of facial landmarks detected using dlib toolbox.

Body features extraction and tracking For body feature, we applied MediaPipe Holistic pipeline [4]; an open-source framework designed with optimized hand tracking, human pose, and face landmark models to generate 33 pose landmarks, 468 face landmarks, and 21 landmarks per hand. For body feature extraction, we considered only the pose and hand landmarks where each landmark consists of x, y coordinates normalized by the image width and height, z that represents the depth of the landmark, and *visibility* which indicates the likelihood of the landmark being visible. The total number of Body features is 300. Figure 5.6 presents examples of pose, right hand, and left hand landmarks for various affect states.

After extracting both face and body features, we apply a softmax function to normalize the data in order to be in the unit simplex.

5.8.3 Affect recognition

In this subsection, we describe the complete affect recognition approach based on Latent SBL model. We elaborate in Figure 5.7 all the steps of affect recognition including: video pre-processing, extracting apex frames, features quantization, features-level fusion, and affect states clustering using Latent SBL where the latent topics are considered as the affect states. We consider for our experiments different number of subjects, affect states, and apex frames. Table 5.3 shows the evaluation

[2] https://pypi.org/project/face-recognition/

[3] http://dlib.net/

[4] https://google.github.io/mediapipe/solutions/holistic

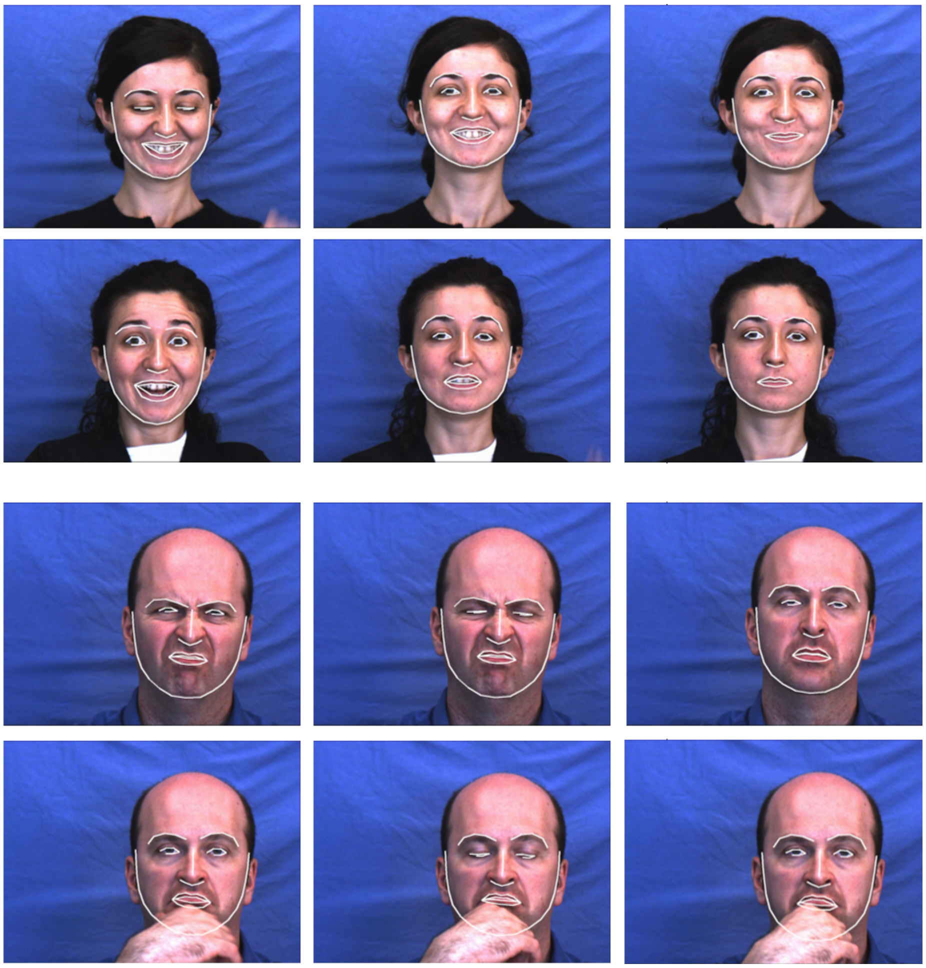


Figure 5.5: Examples of facial landmarks detected using dlib toolbox from faces for different emotions. First row: Happiness. Second row: Positive Surprise. Third row: Anger. Fourth row: Puzzlement.

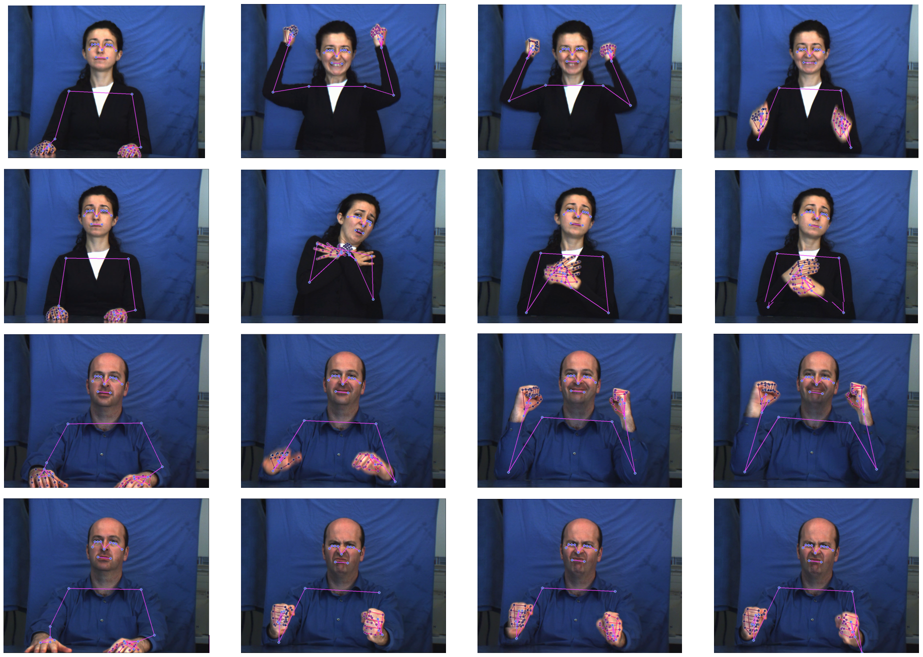


Figure 5.6: MediaPipe Holsitic results that identify human pose, right hand, and left hand landmarks on four affect states: happiness, fear, positive surprise, and anger for two different subjects.

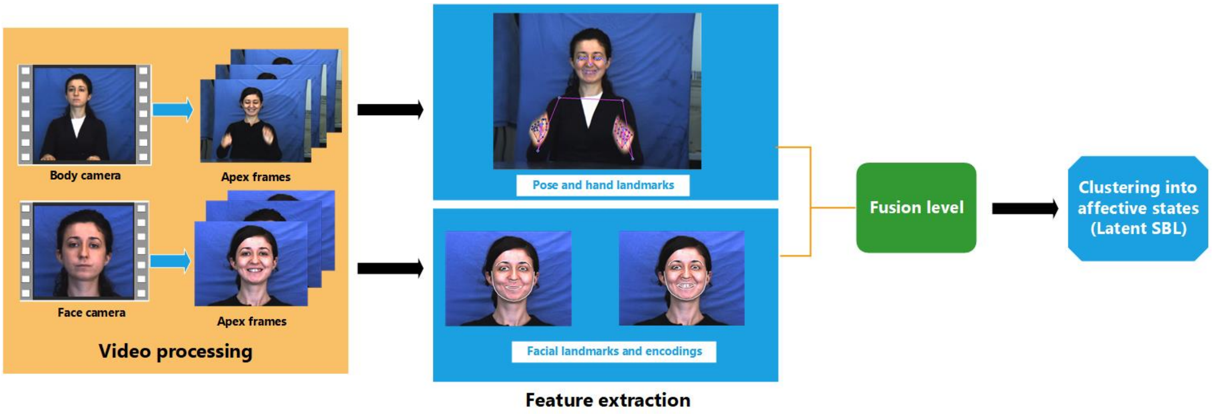


Figure 5.7: The pipeline of affect recognition model from facial-body expressions.

of frame-level affect recognition from face and body features when varying these different metrics. We notice in Table 5.3 the variation of the recognition rate when the number of frames and affect states increase. We mention that Latent SBL model has been proved to be able to detect affect states when the number of classes is higher as well as smaller. The proposed model has been evaluated when taking into account the facial landmarks, the hand gestures, and the body pose features. We observe that the model performs better using hand gestures and pose features which proves that body features provide better information than facial one. Given the fact that both face and body features are high dimensional data with 128 and 300 dimensions, respectively, Latent SBL was able to achieve a recognition rate of 49.63%, 54.54%, and 53.81% using face, hands, and pose features respectively for only 630 apex frames. Detailed accuracy for different frames numbers and affect states is provided in Table 5.3.

Table 5.3: Results of frame-level affect recognition from Face and Body features

Facial landmarks					
Subjects	2	3	6	8	10
Affect states	6	10	12	12	12
Apex Frames	650	1000	2000	3000	4000
Recognition rate (%)	49.63	37.84	31.59	41.20	32.17
Hand gestures					
Subjects	2	3	6	8	10
Affect states	6	10	12	12	12
Apex Frames	650	1000	2000	3000	4000
Recognition rate (%)	54.54	33.50	30.40	30.56	32.8
Body Pose features					
Subjects	2	3	6	8	10
Affect states	6	10	12	12	12
Apex Frames	650	1000	2000	3000	4000
Recognition rate (%)	53.81	31.40	30.45	31.70	30.09

We illustrate the comparison of Latent SBL with some of the related-works that are proposed for FABO system [175] in Table 5.4. We show in Table 5.4 that Latent SBL outperforms existing classification methods and smoothed-based models for the recognition of 12 affective states from face frames. In Table 5.5, it is demonstrated that recognition using concatenated features vectors from face and body improve significantly the results (67.41%) which confirms what has been reported in the state-of-the-art about bimodal affect recognition. To illustrate more the impact of combining the facial landmarks and body features, we present in Table 5.6 qualitative results of monomodal (face and body) and bimodal recognition for video 01 from subject $S001$. We display 9 apex frames of actual label “Happiness” for both face and body modalities. From Table 5.6, it can be noted that almost half of the frames are assigned incorrectly as “Positive surprise” for face modality and ”Fear” using body modality while by feature-level fusion all the frames are perfectly recognized as the actual affect state. We compare also the performance of Latent SBL with related smoothed-based models (SGD, SD) and other standard clustering methods. Overall, our experimental results show that our proposed approach outperforms all the clustering algorithms where the significance of the latent topics and the Beta-Liouville distribution is confirmed. It is noteworthy to mention that the referred related works in Table 5.5 present bimodal classification results from video basis that does not make the evaluation fully comparable. And even that, Latent SBL presents a relative performance with regards to the results of FABO system mentioned in Table 5.5.

Table 5.4: Monomodal recognition results of 12 affective states from face frames using different related classification and clustering methods.

Classifier	Recognized face (%)
BayesNet [175]	28.97
SVM-SOM [175]	32.49
Random Forest [175]	33.56
Adaboost [175]	35.22
Clustering	
SD	26.10
SGD	32.87
Latent SBL	41.20

5.9 Conclusion

In this chapter, we proposed two latent topic modeling approaches based on Beta-Liouville distribution for emotion analysis and affect recognition. First, we presented a new distribution on smoothed simplex where we addressed the challenges of proportional data. Second, we proposed a Latent SBL from which we represented topic mixtures by SBL distributions and we estimated the

Table 5.5: Bimodal recognition results (12 affect states) of the combined Face and Body features using different related classification and clustering methods.

Classifier	Recognized rate (%)
BayesNet [175]	72.73
Random Forest [175]	80.72
Neural Networks[175]	80.27
Adaboost [175]	82.65
Clustering	
K-means	42.29
Affinity propagation	22.16
Agglomerative clustering	42.29
Mean shift	53.77
GMM	42.22
SD	47.79
SGD	54.71
Latent SBL	67.41

Table 5.6: Qualitative results of Monomodal recognition (Face, Body) and Bimodal Fusion-level for S001 Video 01

Apex Frames	Actual affect state	Face modality	Body modality	Bimodal fusion
38	Happiness	Happiness	Happiness	Happiness
39	Happiness	Happiness	Happiness	Happiness
40	Happiness	PST-Surprise	Happiness	Happiness
41	Happiness	Happiness	Happiness	Happiness
42	Happiness	Happiness	Happiness	Happiness
43	Happiness	PST-Surprise	Fear	Happiness
44	Happiness	Happiness	Fear	Happiness
45	Happiness	PST-Surprise	Fear	Happiness
46	Happiness	PST-Surprise	Fear	Happiness

model parameters using a maximum likelihood approach. Third, we introduced a new Emotion-Term model based on Bayesian folding-in with SBL kernels for PLSI. We investigated two challenging applications namely tracking emotions in storytelling scenarios and recognition of affect states from face and body information. For text-based emotion recognition, experimental results proved that the smoothed distribution joined with topics modeling principle leads to a substantially greater ability of tackling textual data challenges. Additionally, the merits of generating SBL kernels for PLSI and defining a new Bayesian folding-in approach have been shown through estimating new unknown stories for the purpose of tracking emotions in fairy tales. On the other hand, Latent SBL successfully recognized affect states from face and body information within a bimodal framework. Results revealed that the novel affect recognition framework mostly outperforms the related smoothed-based models and the standard clustering algorithms and achieves comparable performance to the supervised models applied on FABO database. Taking into consideration that we compared our experimental frame-based results to the related-works which consider classifying affect states from video basis, we aim to improve our framework to be able to track affect states from video sequences. Regarding the promising results of Latent SBL and SBL-ETM, we advocate seeking to analyze emotion states from other modalities such as voice signals and video sequences. Finally, it seems worth exploring SBL models for other interesting applications in biological sciences, business management, and also in medicine.

Chapter 6

On Smoothing and Scaling Language Model for Sentiment Based Information Retrieval

Sentiment analysis or opinion mining refers to the discovery of sentiment information within textual documents, tweets, or review posts. This field has emerged with the social media outgrowth which becomes of great interest for several applications such as marketing, tourism, and business. In this work, we approach Twitter sentiment analysis through a novel framework that addresses simultaneously the problems of text representation such as sparseness and high-dimensionality. We propose an information retrieval probabilistic model based on a new distribution namely the Smoothed Scaled Dirichlet distribution. We present a likelihood learning method for estimating the parameters of the distribution and we propose a feature generation from the information retrieval system. We apply the proposed approach Smoothed Scaled Relevance Model on four Twitter sentiment datasets: STD, STS-Gold, SemEval14, and SentiStrength. We evaluate the performance of the offered solution with a comparison against the baseline models and the related-works.

6.1 Introduction

Information retrieval (IR) plays a dominant role in web-based search engines [187]. Broadly, the field of IR deals with more domains including natural language processing (NLP) [188], image and video retrieval [189], [190], recommendation systems [191], handwriting retrieval [192], and question answering [193]. Recently, IR approaches were oriented successfully to sentiment analysis such as Sentiment Analysis Based on Information Retrieval (SABIR) [194] and Sentiment Analysis for Pseudo-Relevance Feedback [195]. In fact, IR deals with extracting relevant information within large collection of documents. Particularly, Ad-hoc retrieval is a text-based retrieval which responds to a specified user's request/query by retrieving information in the form of output documents that are mostly relevant. Matching the relevant documents to queries requires ranking functions that are commonly based on similarity measures. Sentiment analysis is extensively known as polarity classification of documents according to a particular opinion. The majority of text-based sentiment analysis are based on tweets to classify them into positive or negative sentiment polarity. Sentiment analysis based on information retrieval classify documents or tweets using information about the similarity of unlabeled tweet considered as a query in relation to the elements of labeled tweets.

Existing algorithms addressed for binary sentiment analysis are generally based on machine learning such as the supervised learning techniques: Naive-Bayes, Maximum Entropy, and Support Vector Machines, [196]. Most of the standard models consider using bag-of-words (BoW) for representing the words in tweets as features such as n-grams [197], [198]. Advanced studies present different sets of features that bring new challenges to sentiment analysis including emoticons, hashtags, retweets, and emojis. Recently, deep learning algorithms gained interest for several applications and in particular sentiment analysis. For instance, a convolution neural network was presented in [199] to classify sentiment of movie reviewers using dynamic k-Max pooling. In [200], authors proposed the use of feature hashing to address the problem of sparsity in tweets when using the bag-of-words approach. Jianqiang et al. introduced [201] the combination of deep neural network and word embedding approaches, GloVe-Deep Convolution Neural Network (DCNN) to predict sentiments based on GloVe, a deep word embedding algorithm which represent sentiments information. More recently, feature extraction task has been performed using information retrieval systems to derive sentiments from the tweets. Kauer et al. [194] proposed Sentiment Analysis Based on Information Retrieval (SABIR) that employs features derived from ranking function generated by an information retrieval system in response to a query. Taking into account the intuition behind considering the domain of such document to decode sentiment polarity, a multi-domain sentiment analysis approach was proposed by applying information retrieval techniques for representing information about the linguistic structure of sentences [202]. In a book search domain, Htait et al. [195] proposed a sentiment analysis based query expansion. They introduced a sentiment oriented method for the purpose of selecting sentences from the reviews of top rated book.

In this work, we are mainly concerned with sentiment analysis based on probabilistic approaches for information retrieval that are based on document likelihood. Likelihood-based generative models provide a meaningful function on unseen data and modeling features from documents in the collection. The likelihood function is a successful way to learn from unlabelled data and to sample documents from a particular distribution that best suits the data. We propose a Smoothed Scaled information retrieval approach namely a Smoothed Scaled Dirichlet Relevance Model (SSD-RM). The proposed information retrieval model is based on a new distribution so-called Smoothed Scaled Dirichlet (SSD) that we present in this paper. Specifically, we introduce the maximum likelihood estimation of the parameters that will be used in the retrieval framework.

We review the existing retrieval models in Section 2 where we present the difference between deep matching models and the probabilistic ones. We introduce the challenges faced by probabilistic models and in particular the likelihood-based generative models. Based on that, we provide the motivations that led us to propose the new sentiment-based information retrieval approach. In Section 3, we introduce the Smoothed Scaled Dirichlet prior, the novel distribution, and the maximum-likelihood estimation of SSD parameters. In Section 4, we present the sentiment analysis framework based on SSD. We evaluate the proposed framework in Section 5 on several benchmarks of sentiment analysis, namely, Stanford Twitter sentiment (STD) corpus, Stanford sentiment gold standard (STS-Gold), SemEval2014 Task9, and Sentiment Strength Twitter (SentiStrength) where we compare the classification results with the baselines and state-of-the-art (SOTA) models in the literature. We conclude this work with future directions in Section 6.

6.2 Related work

In the past decades, several information retrieval methods have been proposed [203], [204] which can be categorized into two types according to the model architecture namely deep matching models and probabilistic approaches. We review briefly the relevant approaches applied to information retrieval and we illustrate the motivations of this work in the following.

6.2.1 Deep matching models

Deep learning models applied to information retrieval have formalized the task as a matching problem between documents and queries [205], [206], [207], [208]. When dealing with matching structured objects in natural language, a convolutional matching model namely ARC-I has been proposed for matching two sentences [209]. The ARC-I is based on deep convolutional network and adopts the hierarchical structure of sentences through layer composition and pooling. A semantic modeling using deep auto-encoders integrates a semantic structure when representing documents and the query in latent semantic space [210]. An extended work in [205] makes use of Deep Neural Network (DNN) to propose a series of Deep Structured Semantic Models (DSSM) for web searching. Embedding the notion of such *interestingness* in the semantic model [211], [212] extends the DSSM through a Convolutional Neural network (CNN).

6.2.2 Probabilistic models

Earlier work on information retrieval has been based on the probability ranking principle developed by Robertson [213] in which documents D are ranked by the probability of belonging to a relevant class: $p(D|R)/p(D|N)$. We refer to R as the class of relevant documents and N the class of documents that are non-relevant to the user's query. Typically, the majority of information retrieval technologies seek to rank documents based on supervised methods where the collections is classified into relevant and non-relevant classes of documents. The estimation of $p(D|R)$ differs in various approaches where the probability of words could be sampled through multiple Bernoulli distributions [214], a mixture of multinomial distributions [214] known as unigram models, or with Dirichlet smoothing as in [81]. These IR approaches are described as likelihood-based generative models where the ranking function uses the likelihood of training data.

Recent probabilistic models of information retrieval consider separately documents as models and queries as fixed sample of text generated from these models. The generation is not necessary sampled from the same document model but there is also a possibility to be sampled from a particular relevance model. Besides, topic models have been introduced as probabilistic models that generate documents from a mixture of topics. Topic models gained a lot of attention in document analysis and, in particular, information retrieval. We mention the well-known Latent Dirichlet Allocation (LDA)-based information retrieval [188], [215]. The matching problem has been formalized not only using probabilistic models and deep models [206], [207], [216] but also graphs which have been successfully applied as a language model for information retrieval applications [217].

6.2.3 Motivations

The related-works of IR approaches based on probabilistic models in the literature have used only Bernoulli and multinomial distributions. In fact, these models are very basic to deal with the challenges of short texts and the sparsity problem. Further, a Smoothed Dirichlet distribution have been considered for information retrieval in the work of Nallapati [89] where he considered the probability ranking principle. In this work, he proposed two models for IR, one as relevance model and the second as Smoothed Dirichlet based classification where the cross-entropy was considered as ranking function. Motivated by the promising results obtained by these approaches that demonstrate the capability to capture term occurrence patterns in documents better than the multinomial and taking into account the known limitations of the Dirichlet distribution, we propose at first a new scaled distribution defined in a smoothed simplex. Then, based on this novel distribution, we propose a new probabilistic model for IR.

6.3 Smoothed Scaled Information Retrieval

6.3.1 Scaled Dirichlet prior

In this section, we present a new smoothing approach using a Scaled Dirichlet (SD) prior. In fact, when representing a document $D = (w_1, \ldots, w_V)$ with a Dirichlet distribution, $w_v = \frac{z_v}{\sum_d z_v}$ where each z_v is a standard Gamma-distributed random variable with shape parameter α_v and equally scaled, i.e., $z_v \sim Ga(\alpha_v, 1)$. In fact, the Dirichlet distribution has several drawbacks such as the negative covariance structure, the estimation of the parameters is not straightforward and the MLE has no simple closed form solution which leads to computationally expensive learning algorithms. In this regard, the SD distribution [218] is more flexible than the Dirichlet distribution since it removes the requirement of equally scale parameter of Gamma variables i.e. $z_d \sim Ga(\alpha_d, \beta_d)$. In fact, the scaled Dirichlet distribution is obtained from a perturbed composition with a Dirichlet distribution which defines a vector-space structure in the simplex. The proposed language model is a multinomial distribution for which the conjugate prior for Bayesian analysis is a Scaled Dirichlet distribution with shape and scaling parameters, $(\alpha_1, \ldots, \alpha_V), (\beta_1, \ldots, \beta_V)$ respectively. Thus, the probability density function assigned to each document D is given by:

$$p(w_1, \ldots, w_V | \vec{\alpha}, \vec{\beta}) = \frac{\Gamma(\alpha_+)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \frac{\prod_{v=1}^{V} \beta_v^{\alpha_v} w_v^{\alpha_v - 1}}{(\sum_{v=1}^{V} \beta_v w_v)^{\alpha_+}}. \tag{142}$$

where $\alpha_+ = \sum_{v=1}^{V} \alpha_v$ and $\Gamma(.)$ denotes the Gamma function. Note that SD is a generalization that reduces to Dirichlet distribution when all the scaled parameters are all equal $\beta_1 = \cdots = \beta_D = \beta$.

6.3.2 Smoothing Scaled Dirichlet simplex

Given a count representation from a proportion vector f_v; normalized as $w_v^D = \frac{f_v}{|D|}$ where $|D|$ represents the length of document D, the compressed domain that contains all the smoothed language models is given by the following:

$$\Delta^s = \{\lambda w_v^D + (1 - \lambda) w_v^C; w_v^D \in \Delta\}. \tag{143}$$

where $0 < \lambda < 1$ is a smoothing parameter used to smooth a document language model with $w_v^C = \frac{\sum_{i=1}^{C} f_{iv}}{|C|}$, *i.e.* the proportion of words occurring in the collection C of all documents considered.

Estimating the smoothed language model under the Scaled Dirichlet distribution that considers the whole simplex $\Delta = \{w | \forall_d w_d > 0; \sum_{v=1}^{V} w_v = 1\}$ will incorrectly represents the collection since documents span only a small fraction of the entire domain (this phenomenon is illustrated with a detailed example in [89]). Hence, one way to overcome this problem is to propose a new domain spanned by the smoothed documents. This can be illustrated by presenting a new variation of the Scaled Dirichlet distribution called the Smoothed Scaled Dirichlet (SSD) distribution with proposing an approximated normalizer over the compressed domain. We start by giving the Scaled Dirichlet normalizer which is defined by:

$$Z^{SD} = \int_{\vec{w}\in\Delta} \frac{\prod_{v=1}^{V} w_v^{\alpha_v - 1}}{(\sum_{v=1}^{V} \beta_v w_v)^{\alpha_+}} d\vec{w} = \frac{\prod_{v=1}^{V} \Gamma(\alpha_v)}{\Gamma(\alpha_+) \prod_{v=1}^{V} \beta_v^{\alpha_v}}. \tag{144}$$

Taking into account the smoothed representation of documents, the integral of the normalizer should only be spanned over the compressed domain Δ^s, which gives:

$$Z^{SSD} = \int_{\vec{w}\in\Delta^s} \frac{\prod_{v=1}^{V} (w_v^s)^{\alpha_v - 1}}{(\sum_{v=1}^{V} \beta_v w_v^s)^{\alpha_+}} d\vec{w}^s. \tag{145}$$

Exploiting the mapping from Δ^s to Δ of the smoothing language model defined in equation 143 into 145, we get:

$$Z^{SSD} = \lambda \int_{\vec{w}\in\Delta} \frac{\prod_{v=1}^{V} (\lambda w_v^D + (1-\lambda) w_v^C)^{\alpha_v - 1}}{(\sum_{v=1}^{V} \beta_v (\lambda w_v^D + (1-\lambda) w_v^C))^{\alpha_+}} d\vec{w}. \tag{146}$$

We note that SSD distribution has the same parametric form as the Scaled Dirichlet distribution but with a different normalizer defined over the smoothed simplex which gives:

$$p(w_1^s, \ldots, w_V^s | \vec{\alpha}, \vec{\beta}) = \frac{1}{Z^{SSD}} \frac{\prod_{v=1}^{V} (w_v^s)^{\alpha_v - 1}}{(\sum_{v=1}^{V} \beta_v w_v^s)^{\alpha_+}}. \tag{147}$$

Inspired by the approximation of the smoothed Dirichlet normalizer [82] where the concept is to develop an analytically tractable solution using Stirling' approximation for the Gamma function, we define the following:

$$\begin{aligned} Z^{SSD} &= \frac{\prod_{v=1}^{V} \Gamma_a(\alpha_v)}{\Gamma_a(\alpha_+) \prod_{v=1}^{V} \beta_v^{\alpha_v}} \simeq \frac{\prod_{v=1}^{V} e^{-\alpha_v} \alpha_v^{\alpha_v}}{e^{-\sum_{v=1}^{V} \alpha_v} (\alpha_+)^{\alpha_+} \prod_{v=1}^{V} \beta_v^{\alpha_v}} \\ &\simeq \frac{\prod_{v=1}^{V} \alpha_v^{\alpha_v}}{(\alpha_+)^{\alpha_+} \prod_{v=1}^{V} \beta_v^{\alpha_v}}. \end{aligned} \tag{148}$$

where $\Gamma_a(\alpha) \simeq e^{-\alpha}\alpha^{\alpha}$ is the approximation of the Gamma function that makes sure the unboundedness of the normalization and yields a closed form solution to the maximum likelihood estimation. Thus, we define the Smoothed Scaled Dirichlet distribution as follows:

$$p(w_1^s, \dots, w_V^s|\vec{\alpha}, \vec{\beta}) = \frac{(\alpha_+)^{\alpha_+} \prod_{v=1}^V \beta_v^{\alpha_v}}{\prod_{v=1}^V \alpha_v^{\alpha_v}} \frac{\prod_{v=1}^V (w_v^s)^{\alpha_v - 1}}{(\sum_{v=1}^V \beta_v w_v^s)^{\alpha_+}}. \tag{149}$$

6.3.3 Maximum-Likelihood estimation

Given a set of N documents $\mathcal{D} = (D_1, \dots, D_N)$ over K number of clusters, we define the log-likelihood function corresponding to SSD distribution as follows:

$$\mathcal{L}(\mathcal{D}|\vec{\alpha}, \vec{\beta}) = \sum_{i=1}^{N} \log \Big(p(w_{i1}^s, \dots, w_{iV}^s|\vec{\alpha}, \vec{\beta}) \Big). \tag{150}$$

Differentiating the log-likelihood with respect to each α_v and β_v (appendix B) with treating α_+ as a constant and equating to zero, we get the following Maximum-Likelihood estimates (MLE) of α_v and β_v:

$$\hat{\alpha}_v = \sum_{i=1}^{N} \beta_v \frac{w_{iv}^s}{\sum_{v=1}^V \beta_v w_{iv}^s}. \tag{151}$$

$$\hat{\beta}_v = \sum_{i=1}^{N} \frac{\alpha_v}{\alpha_+} \frac{\sum_{v=1}^V \beta_v w_{iv}^s}{w_{iv}^s}. \tag{152}$$

The MLE of the parameters of SSD provides closed-form solutions.

6.3.4 SSD-based retrieval

In this section, we propose a new information retrieval approach so-called Smoothed Scaled Dirichlet Relevance Model (SSD-RM). SSD-RM is defined as a relevance model where we calculate the Kulback-leibler divergence between the query and each document to select the n top ranking documents measured by query-likelihood [219]. Given a query $Q = (q_1, \dots, q_d)$, the relevance model $\mathcal{P}_R$ based on a language model framework, is defined from the expected value of top ranking documents as:

$$\begin{aligned} \mathcal{P}_R &= E[\mathcal{D}|Q] = \sum_{i=1}^{n} D_i p(D_i|Q) \\ &= \sum_{i=1}^{n} D_i \frac{p(Q|D_i)p(D_i)}{\sum_{i=1}^{n} p(Q|D_i)p(D_i)}. \end{aligned} \tag{153}$$

We assume a uniform prior over all the documents $p(D_i)$. Thus, the computation of the relevance model reduces to weight average of the query-likelihood $p(Q|D_i)$ that determines how well a document D_i fits the query Q:

$$p(Q|D_i) = \frac{p(D_i|Q)p(Q)}{p(D)}. \tag{154}$$

Taking into account that all documents have the same prior and considering the fact that, when a query is on the same topic as the document, the language model will induce a high query score approximated using the multinomial distribution as per the principle of the known relevance model in the SOTA [220]. Hence, we define the likelihood to generate a Relevance Model (RM) as:

$$\begin{aligned} \log p(Q|D_i) &= \sum_{v=1}^{V} q_v \log(w_{iv}^s) \\ &= \sum_{v=1}^{V} q_v \log\Big((1-\lambda)w_i^C\Big[\frac{\lambda w_{iv}^D}{(1-\lambda)w_{iv}^C} + 1\Big]\Big) \\ &\propto \sum_{v=1}^{V} q_v \log\Big(\frac{\lambda w_{iv}^D}{(1-\lambda)w_{iv}^C} + 1\Big). \end{aligned} \tag{155}$$

By defining a scaled prior over the smoothed simplex using Multinomial Scaled Dirichlet distribution [103], we get:

$$
\begin{aligned}
\log p(Q|D_i) &\propto \sum_{v=1}^{V} \log\Gamma(q_v + w_{iv}^s) - \log\Gamma(q_v) && (156)\\
&\overset{\Gamma_a}{\propto} \sum_{v=1}^{V} (q_v + w_{iv}^s)\log(q_v + w_{iv}^s) - q_v \log(q_v) - w_{iv}^s \\
&\propto \sum_{v=1}^{V} q_v \Big(\log(q_v + w_{iv}^s) - \log(q_v) \Big) \\
&\propto \sum_{v=1}^{V} q_v \log\Big(\frac{(1-\lambda)w_{iv}^C + \lambda w_{iv}^D}{q_v} + 1 \Big).
\end{aligned}
$$

where $\overset{\Gamma_a}{\propto}$ refers to the approximation $\Gamma_a(\alpha) \simeq e^{-\alpha}\alpha^{\alpha}$ used to define the Smoothed Scaled distribution. Thus, the weight average of the Smoothed Scaled Relevance model is normalized over the top ranking documents to give:

$$
\mathcal{P}_R = \frac{1}{Z} \sum_{i=1}^{n} D_i \exp\Big\{ \sum_v q_v \log\Big(\frac{(1-\lambda)w_{iv}^C + \lambda w_{iv}^D}{q_v} + 1 \Big) \Big\}. \qquad (157)
$$

where Z is the normalizer that sums up the average weight over all the top ranking documents. Using the expected value assigned by SSD relevance model $\mathcal{P}^R$, we define the score calculated between a selected query and a ranking document over the same vocabulary V using the Kulback-Leibler divergence:

$$
\begin{aligned}
\text{Score}(Q, D) &= -KL(Q||D) && (158)\\
&= \sum_{w \in V} p_q(w) \log \frac{p_q(w)}{p_d(w)}.
\end{aligned}
$$

where the score measures the relative entropy between the two distributions $p_q(w)$ and $p_d(w)$, where $p_q(w)$ encodes the relevance model $\mathcal{P}_R$ and $p_d(w)$ the SSD document model over the parameter space (α, β) computed according to equation 149.

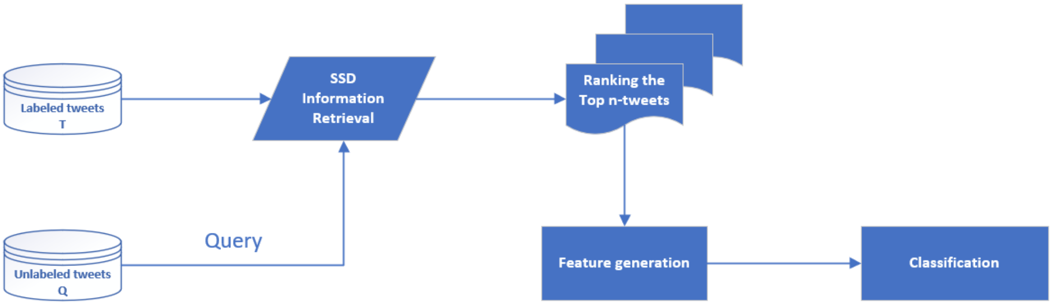


Figure 6.1: Overview of sentiment based SSD-RM

6.4 Sentiment Based Information Retrieval

In this section, we describe our proposed sentiment analysis framework based on Smoothed Scaled Dirichlet relevance model. Our framework is composed of three main steps: information retrieval system, feature generation, and classification. We present an overview of the process in Figure 6.1.

6.4.1 Information retrieval system

In response to a query set, information retrieval algorithms are applied to rank the top similar documents. In this work, we consider the set of documents as labeled tweets and rank them in relation to each unlabeled tweet taken as a query. Given a set of tweets, we split the dataset into two subsets: labeled tweets for training $T = \{t_1, \ldots, t_m\}$ for which the class is known (positive or negative) and unlabeled tweets $Q = \{q_1, \ldots, q_d\}$ where each tweet is considered as query. To select the n-top tweets using our method, the first step is to index labeled tweets using the proposed SSD-RM. Second, we rank the n most similar tweets using the score function defined by Kulback-Leibler (Eq. 159).

6.4.2 Feature generation

In this subsection, we explain how to extract 24 features (12 for each class) from top ranking tweets. We extract the same features suggested in the work of [194]. First, we determine the aggregation functions for each class such as max, min, sum, average, and count. Second, we derive a new feature ϕ that takes into consideration the absolute and the relative ranks of each tweet related to each class as follows:

$$\phi_c = \sum_{r=1}^{n_c} \frac{rank_{rel}}{rank_{abs}} \tag{159}$$

where c is the class (positive or negative), n_c is the number of ranked tweets for that class, $rank_{rel}$ is the relative position of the tweet regarding the tweets of that class, and $rank_{abs}$ is the absolute position of the tweet in the overall ranking.

Following, we then derive other features from the combination of ϕ with the aggregation functions as follows: [194] $\phi_{max_c} = \frac{\phi_c}{max_c}$, $\phi_{min_c} = \frac{\phi_c}{min_c}$, $\phi_{sum_c} = \frac{\phi_c}{sum_c}$, $\phi_{avg_c} = \frac{\phi_c}{avg_c}$, and $\phi_{count_c} = \frac{\phi_c}{count_c}$ where max_c, min_c, are the maximum and minimum scores for class c, sum_c, avg_c, compute the sum and average scores respectively, and $count_c$ determines the number of tweets from class c in the ranking. The last feature combines the scores of the retrieved tweets and their positions in the ranking, given by:

$$\phi_{positional_c} = \sum_{r=1}^{n} \frac{rank_{rel}}{rank_{abs}} \text{ score}(q_i) \tag{160}$$

We generate new features from IR system instead of the bag-of-words structure to tackle the problem of dimensionality. In fact, we present usually tweets using the BOW structure in the format of high-dimensional feature vectors which provoke more the sparsity issue. However, taking only a 24-feature vector, makes the classification task easier and it improves performance as we are able to classify tweets with low-dimension. In addition, features extracted from IR system reflect the similarity between tweets and lexical properties of words in the whole dictionary.

6.4.3 Classification

The last step of our proposed framework is to classify the test set of tweets. After deriving the new features from the retrieval system, we classify them using a supervised learning algorithm. We tested different machine learning algorithms (SVM, K-NN, and Logistic Regression) to select the one that gives the best classification performance.

6.5 Experiments analysis

In this section, we evaluate our method on four different balanced datasets for sentiment analysis: Stanford Twitter sentiment (STD), Stanford Sentiment Gold Standard (STS-Gold), SemEval2014 Task9, and Sentiment Strength Twitter (SentiStrength). We selected these datasets for three reasons: (1) they are publicly available to theresearch community, (2) they have been extensively used in the literature for Twitter sentiment classification, and they are (3) manually annotated as positive or negative sentiments labels. We consider in the ad-hod IR system experiments a binary classification. Before applying the IR system on the labeled and unlabeled tweets, we pre-process the datasets and present the tweets using TF-IDF approach. Next, as we tested different machine learning algorithms (SVM, K-NN, and Logistic Regression), we found that K-NN gives the best performance among the others. All the listed results in the next sections are obtained using the K-NN algorithm. We compare the obtained performance of our model with different models including the baseline machine learning algorithms (SVM, KNN, MNB, GNB) and SOTA approaches applied on these datasets. Furthermore, we study the influence of the number of ranked tweets and the dimension of features on the performance of the proposed framework.

6.5.1 Datasets

- **Stanford Twitter sentiment (STD)** [1] known also as Sentiment140 dataset introduced by Go et al. [221]. The STD corpus contains STD-train and STD-test subsets which contain 1,6 million tweets labeled into positive, negative, and neutral sentiment tweets. The STD-test set consists only of 182 positive, 177 negative, and 139 neutral tweets and STD-train consists of 248 576 positive and 800 000 negative tweets. For our experiments, we extract a subset of tweets from STD-train containing only 8000 tweets equally divided into positive and negative.

- **Stanford Sentiment Gold Standard (STS-Gold)** [2] dataset constructed by Saif et al. [222] which is extracted from the original Stanford Twitter corpus using AlchemyAPI[3] online service. It contains 2034 positive and negative tweets annotated manually by three graduate students.

[1] Available at http://help.sentiment140.com/

[2] Available at https://www.kaggle.com/divyansh22/stsgold-dataset/

[3] http://www.alchemyapi.com/

This dataset takes into account the associated semantic concept class such as the city, person, etc.

- **SemEval2014 Task9 (SemEval14)** [4] provided in the Semantic Evaluation of Systems challenge SemEval-2014 for the Twitter sentiment analysis task (task 9) [223]. The original SemEval dataset contains three subsets: training, development and test with a total of 20K tweets. The dataset was annotated manually by Amazon Mechanical Turk workers into positive, negative, and neutral labels. For applying our framework, we removed the neutral labels to have in total 3263 positive and 1351 negative sentiment labels.

- **Sentiment Strength Twitter (SentiStrength)** [5] introduced in [224] to evaluate SentiStrength for sentiment strength detection on social web. The dataset contains 4242 tweets manually annotated in positive and negative labels where the positive strength range between 1 (not positive) and 5 (extremely positive) and the negative strength is a number from -1 (not negative) to -5 (extremely negative). In this work, we re-annotate the labels following the methodology proposed in [222]. Thus, we assign only two labels: positive and negative. For that, we consider a tweet as positive if its positive sentiment strength is 1.5 times higher than the negative one, and negative otherwise. We obtain 1,037 negative and 1,252 positive tweets.

6.5.2 Comparison with baseline models

In this subsection, we present a comparison of the proposed SSD-RM with baseline models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Multinomial Naive Bayes (MNB), and Gaussian Naive Bayes. We evaluate these models according to the Accuracy, Precision, Recall, and F1-measure. We display the results obtained for the four datasets (STD, STS-Gold, SemEval14, SentiStrength) in Table 6.1. In all datasets, SSD-RM achieves the best performance. Therefore, in this study, our proposed method was able to increase the performance of sentiment classification up to 80% accuracy for STD, STS-Gold, and SemEval14 datasets, and for SentiStrength corpus around 70% accuracy. In terms of precision, recall, and F1-measure, SSD-RM outperforms the other baseline models for all the datasets except for SemEval14 where SVM

[4] Available at https://alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools

[5] Available at http://sentistrength.wlv.ac.uk/documentation/

achieves a precision of 97.10%. The differences between SSD-RM and the other tested models are statistically significant: using student's t-test, p-values are between 0.025 and 0.018 for the STD dataset, in the range of $[0.0003, 0.035]$ for STS-Gold, and between 0.00057 and 0.00037 for both SemEval14 and SentiStrength datasets.

Table 6.1: Classification results obtained by SSD-RM and baseline models (SVM, KNN, MNB, GNB) for the selected datasets: STD, STS-Gold, SemEval14, SentiStrength.

Dataset		SVM	KNN	MNB	GNB	SSD-RM
STD	Accuracy	84.15	79.70	84.05	84.05	**84.93**
	Recall	61.69	61.42	61.82	61.29	**62.56**
	Precision	96.48	91.72	**96.38**	96.31	96.32
	F1-measure	74.26	73.57	74.33	74.10	**74.58**
STS-Gold	Accuracy	76.41	71.20	64.11	54.17	**80.72**
	Recall	54.35	**62.50**	43.54	46.13	58.56
	Precision	42.64	38.26	41.39	**80.1**	59.66
	F1-measure	47.80	47.46	42.44	**60.9**	58.3
SemEval14	Accuracy	71.50	63.54	69.6	53.83	**81.04**
	Recall	57.60	59.19	57.64	**61.79**	61.08
	Precision	**97.10**	77.72	94.01	60.46	91.6
	F1-measure	72.31	67.22	71.43	61.12	**72.86**
SentiStrength	Accuracy	62.39	58.77	62.50	56.79	**77.23**
	Recall	55.43	55.22	54.33	59.35	**59.51**
	Precision	76.64	69.60	76.46	53.03	**82.83**
	F1-measure	64.33	61.70	63.73	56.01	**69.07**

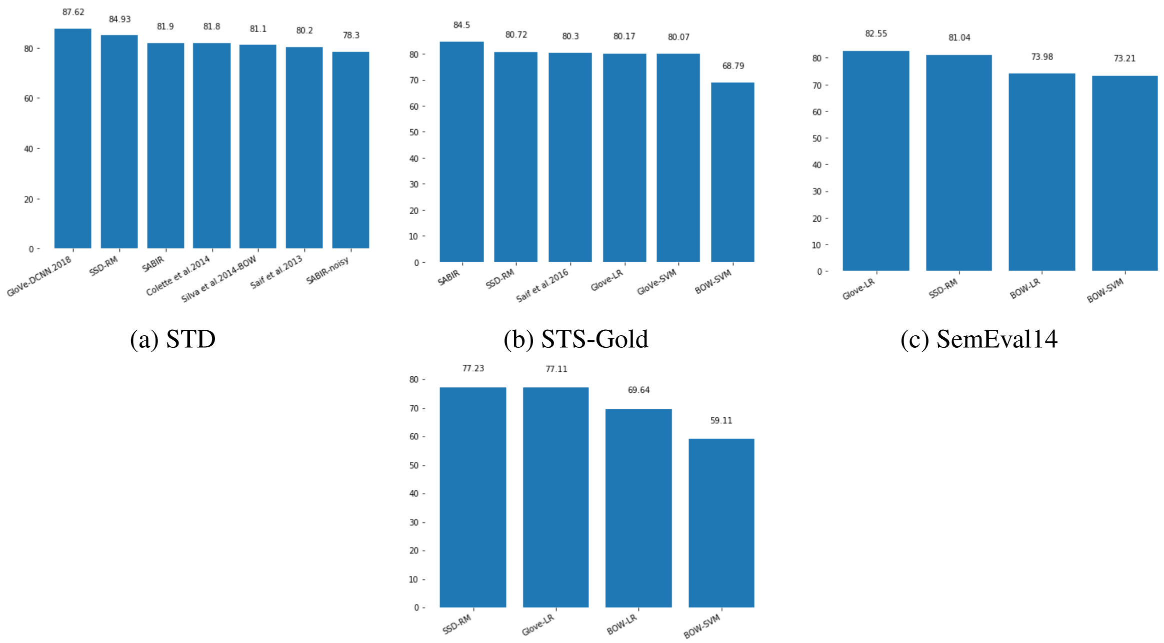
In Table 6.4, we compare our proposed model with other RM models based on multiple Bernoulli [214], multinomial, and smoothed Dirichlet [89] so-called M-RM, B-RM, and SD-RM, respectively. We considered different sparsity rates for the four datasets where sparsity score is defined as the number of zero values in the document collection divided by the total number of words. The classification results demonstrates that SSD-RM outperforms all the other related models with different sparsity rates which shows the importance to consider the Smoothed Scaled distribution for the retrieval model. We show also in Table 6.4, that under five different rates, the proposed IR system outperforms always the other tested methods including the M-RM, B-RM, and SD-RM for the four considered datasets. The sparsity rates are chosen with respect to the vocabulary dimension in the datasets as demonstrated in Table 6.4. We notice that all the four datasets are highly sparse even with low dimension. For example, the sparsity rate for STD datsetset is equals to 97.06 for 50-dimensional features. Thus, our proposed IR approach are able to deal with sparsity issue and offers competitive performance in handling sparse data comparing to the other language models.

6.5.3 Comparison with other published results

We consider different related-works to compare the results with the ones obtained by our proposed framework: SABIR [194], Colette et al. [225], Silva el al. [200], Saif el al. [222], (GloVe-LR, GloVe-SVM, GloVe-DCNN) [201], and (BOW-SVM, BOW-LR) [226]. These approaches can be categorized into two main groups: deep neural networks and machines learning algorithms. Also, the machines learning approaches have considered two different types of feature extractions. The information retrieval methods are based on feature engineering and the other algorithms have used the BOW structure for text representation. We display, in Figure 6.2, the results of accuracy of the different models applied on STD, STS-Gold, SemEval14, and SentiStrength datasets. On the STD and SemEval14 datasets, SSD-RM achieves the second best results where the outperformed approaches GloVe-DCNN and GloVe-LR apply deep convolutional neural network for word representation. In fact, our proposed approach incorporates retrieval information in the framework and feature engineering while the compared models consider co-occurrence matrices. On SentiStrength dataset, our proposed SSD-RM outperforms the other related-works where the two least performing approaches apply BOW structure for word representation. This mention how the IR achieves better in classifying sentiment tweets. For STS-Gold, we found that SSD-RM outperforms both the deep convolutional neural network and also the BOW-SVM method which validate our hypothesis that underline the advantage of constructing the new features from the top ranking tweets. Even so, SSD-RM presents the second best result comparing to the SABIR approach which employs as well an information retrieval system for sentiment classification.

Table 6.2: Classification results comparing the proposed SSD-RM against the retrieval models based on Multinomial (M-RM), Multiple Bernoulli (B-RM), and Smoothed Dirichlet (SD-RM) under different sparsity rates.

Sparsity (%)	M-RM	B-RM	SD-RM	SSD-RM
		STD		
97.06	61.12	70.08	70.01	84.93
97.95	62.62	58.50	66.87	83.15
98.64	69.56	70.75	70.12	82.00
98.93	65.56	70.08	70.25	83.06
99.14	65.25	70.68	70.64	82.18
		TS-Gold		
96.59	60.56	59.82	60.68	76.28
97.56	56.14	60.12	58.10	78.25
98.34	57.01	57.73	61.30	77.03
98.72	57.24	54.54	61.67	78.13
98.96	61.67	55.77	61.42	78.99
		SemEval14		
96.13	60.35	55.25	60.72	77.95
97.42	63.01	63.12	62.01	78.23
98.34	61.15	57.81	61.46	79.08
98.72	63.65	56.60	62.03	79.63
98.93	61.86	63.51	63.03	79.66
		SentiStrength		
96.22	53.62	50.53	48.90	73.90
97.53	53.39	53.72	54.16	78.61
98.42	52.19	53.17	55.48	74.67
98.78	46.27	51.97	52.52	72.47
98.98	50.01	46.60	56.68	77.35



(a) STD

(b) STS-Gold

(c) SemEval14

(d) SentiStrength

Figure 6.2: Accuracy results for the state-of-the-art and SSD-RM

6.5.4 Features performance

We now evaluate the influence of some parameters on the proposed model. We study the performance of SSD-RM in terms of the number of ranking top tweets and the feature dimensions of words. Figure 6.3 displays the accuracy scores of SSD-RM in terms of the number of top-ranked tweets. We varied the number of top retrieved tweets from 10 to 1000 for all the considered datasets. We notice that increasing the number of retrieved tweets gives more classification performance where the best results for the four datasets is around 200. It is clear from Figure 6.3 that after a number of 800 ranked tweets, the performance decreases. This indicates that the more the number of retrieved tweets is important, the less useful the information extracted from the generated features will be useful. Moreover, we evaluate the execution time of our model in terms of feature dimension (see Table 6.3). As we mentioned before in the pre-processing step, we represent words using tf-idf before applying the IR system. We considered features of dimension 50, 100, 200, 500, and 1000. We notice that the higher the feature dimension, the greater is the decrease of accuracy, and the greater the execution time. The best results obtained for STD, STS-Gold, and SentiStrength are for around 100-dimensional features and 50 dimension for SemEval14.

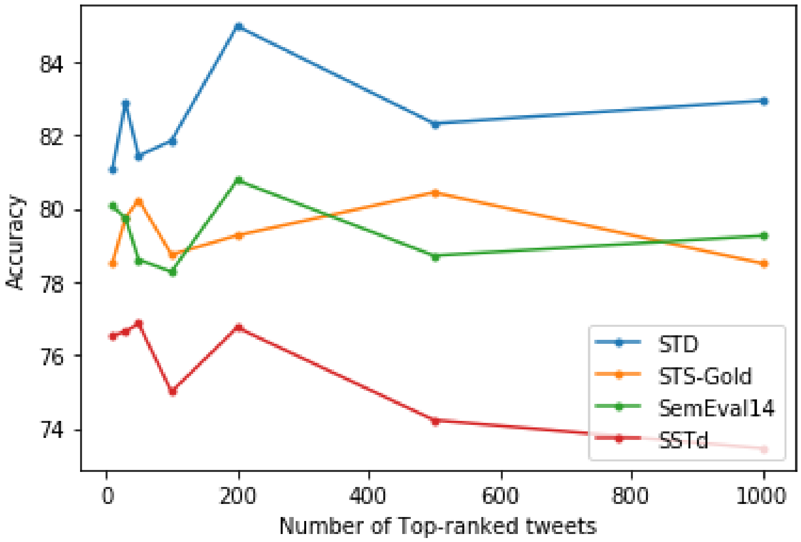


Figure 6.3: Evaluation of SSD-RM in terms of number of ranking top tweets

Table 6.3: Feature influence on time, sparsity, and accuracy for SSD-RM.

Dataset	Features	50	100	200	500	1000
STD						
	Accuracy	84.93	83.15	82	82.37	82.12
	Time (sec)	241.82	250.33	267.27	305.08	403.88
	sparsity (%)	97.06	97.95	98.64	99.25	99.56
STS-Gold						
	Accuracy	80.43	80.72	75.92	79.35	77.48
	Time (sec)	64.68	65.6	73.12	88.12	121.62
	sparsity (%)	96.60	97.50	98.33	99.10	99.45
SemEval14						
	Accuracy	81.04	79.49	79.08	77.98	79.58
	Time (sec)	330	327.42	363.89	419.63	515.3
	sparsity (%)	96.11	97.47	98.33	99.07	99.43
SentiStrength						
	Accuracy	76.86	77.23	77.19	75.76	74.89
	Time (sec)	81.59	84.09	89.64	105.72	153.03
	sparsity (%)	96.22	97.57	98.39	99.15	99.46

Table 6.4: Evaluation of SSD-RM in terms of different sparsity rates using randomly generated sparse data

Dataset	Sparsity rates	25%	50%	75%	90%
STD	Accuracy	96.12	96.18	81.93	80.31
	Recall	65.71	65.83	66.01	65.91
	Precision	92.24	91.83	79.69	79.33
	F1-measure	75.72	76.63	72.21	72.01
STS-Gold	Accuracy	91.52	90.62	77.14	76.65
	Recall	63.78	62.41	57.38	50.01
	Precision	69.84	73.71	54.64	60.25
	F1-measure	66.67	60.51	55.98	54.64
SemEval14	Accuracy	89.81	87.81	78.11	78.71
	Recall	64.27	64.99	59.68	59.85
	Precision	87.28	84.45	93.95	93.31
	F1-measure	74.03	73.45	72.99	72.92
SentiStrength	Accuracy	92.6	91.73	75.21	73.56
	Recall	63.60	63.96	61.37	62.53
	Precision	86.13	84.03	72.91	67.72
	F1-measure	73.23	72.63	66.65	65.02

In Table 6.4, we evaluate the SSD-RM approach in terms of sparsity rates. We mixed a random generated sparse data with each Twitter dataset while varying the sparsity rates and considering the feature dimensions as fixed (1000). It is clearly shown in Table 6.4 that evaluation metrics including accuracy, recall, precision, and F1 measure decrease when sparsity rates are more important. We note the influence of the sparsity on the performance of the retrieval model. Even though the sparsity rates of the data are enormous, the proposed model is able to achieve good performance.

6.6 Conclusion

This chapter presented a new sentiment analysis approach. The proposed approach was based on an IR system and a probabilistic-based model. We first introduced a new distribution defined in a smoothed simplex that was able to handle sparse data namely “Smoothed Scaled Dirichlet”. We proposed an information retrieval approach integrating the SSD and the ranked model principle. The SSD-RM was considered for the purpose of classifying sentiment tweets. Hence, we introduced a sentiment based information retrieval framework that considers a subset of the tweets as queries and generated new features from IR system. The proposed model achieves good results when applying them on four sentiment tweets datasets: STD, STS-Gold, SemEval14, and SentiStrength. We proved the outperformance of our proposed model SSD-RM by comparing the results against the baseline models and the related-works approaches. In a future direction, this work could open different scopes for exploration. Indeed, we can deal with data representation as a first step and integrate instead a deep convolutional neural network approach for that purpose. Further, it is interesting to consider other applications than sentiment analysis and takes into account multi-classification.

Chapter 7

Sparse adaptive Bayesian multinomial estimation using vocabulary knowledge

Popularity of count data is accompanied by its challenging nature such as high-dimensionality and sparsity. Multinomial distribution and extensions are widely applied for modeling data with multivariate count sequences where smoothing the parameters of multinomial distributions is an important concern in statistical inference tasks. This chapter considers the strength of estimating multinomial parameters based on Bayesian methodology using two different hierarchical priors. Respect of this, we propose first a probabilistic approach using a hierarchical generalized Dirichlet prior for sparse multinomial distributions. Our technique builds up on Bayesian knowledge over large discrete domains represented by subsets of feasible outcomes: observed and unobserved. This model allows us to predict the new outcomes based on the preceding data generated from a multinomial distribution. Second, we present another smoothing prior for the Multinomial Naive Bayes classifier which takes advantage of the Beta-Liouville distribution for the estimation of the multinomial parameters. Dealing with sparse documents, we exploit vocabulary knowledge to define two distinct priors over the "observed" and the "unseen" words. We analyze the problem of large-scale and sparse data by enhancing Multinomial Naive Bayes classifier through smoothing the estimation of words with a Beta-scale. We evaluate the sparse generalized multinomial over large benchmarks associated with emotion prediction through two different experiments results in competitive performance. The first experiment reveals predicting emotions in poetry context from English and German dictionaries. The second experiment concerns analyzing flow of emotions related to natural disasters. Next, the novel Beta-Liouville Naive Bayes with vocabulary knowledge is evaluated on two different challenging applications with sparse and large-scale documents namely: emotion intensity analysis and hate speech detection. Experiments on real-world datasets show the effectiveness of our proposed classifier compared to the related-work methods.

This chapter includes two manuscripts and organized as follows. Section 2 introduces the first approach published in International Conference on Advanced Data Mining and Applications. Section 3 provides the second manuscript published in the International Conference on Document Analysis and Recognition. We conclude the two contributions in Section 4.

7.1 Sparse Generalized Dirichlet Prior based Bayesian Multinomial estimation

Data with multivariate count sequences known also as count data abound in many statistical domains including language modeling, political sciences, financial studies, and biology [7, 227, 109, 108]. In this perspective, modeling count data has gained a great attention where the multinomial distribution has been known as the classical model considered for this type of data. However, this distribution was not robust enough to deal with the challenges of count data including sparsity, burstiness, and overdispersion [69]. Improving estimates over multinomial approach leads to incorporate Bayesian knowledge using conjugate priors such as the well-known Dirichlet-compound-multinomial (DCM) [22]. The DCM has been effectively applied in artificial intelligence where it has been considered as a key model in various domains including probabilistic graphical models [228], smoothing methods for language models [81], and topic modeling [229]. While in other line of research, methods address the problem of choosing the Dirichlet prior under a mean-squared error criterion [230, 231] and additional employments of the Dirichlet prior for the multinomial probabilities [232]. Going beyond the difficulties and the weak properties of the Dirichlet prior, more flexible priors have been introduced in [27] where the generalized Dirichlet has been considered to build the multinomial generalized Dirichlet mixture model. Besides, the Beta-Liouville which belongs to the multivariate Liouville family is a conjugate prior to the multinomial distribution [102]. A recent count data modeling approach has adapted the scaled Dirichlet [103] and the shifted-scaled Dirichlet [126] as conjugate priors to propose statistical frameworks for clustering count data.

Despite all the significant contributions to the modeling of count data, all the above mentioned research works have considered smoothing the multinomial parameters and integrating it out to define a new distribution based on the parameters of the conjugate prior. Unfortunately, there are no considerable research works for modeling sparse count data while keeping the estimation of multinomial cell probabilities. With respect to statistical language processing, in the context of parameter estimation for multinomial cell probabilities [233, 234], Bayesian approaches that employ weak prior knowledge generally face the problem of sparseness and high-dimensionality. In this context, the probability of characters was assigned to subsets of different cardinalities (small, large and moderates size of alphabets) using natural law [235]. Estimating sparse multinomial parameters using Dirichlet prior have been considered in [236, 237]. In the work of [237], authors address the problem of large corpora in natural language application when distributions were mostly assigned to words that were not seen in the training set. They propose a Bayesian approach considering a hierarchical Dirichlet prior for multinomial distribution over exponential hypothesis each of which represents a set of feasible outcomes. In [236], the matter of uncertainty in vocabularies was addressed to give more flexibility with regards to the method of Friedman and Singer [237].

Meanwhile, the Dirichlet prior employed in [237] and [236] favors the sparsity due to the definition of proportions in the unit simplex using only one shape parameter. In this regards, sparse outcomes will be assigned low values of shape parameter which leads to prior probabilities located in the corner of the simplex implies to have only small or large probabilities. Consequently, the generalized Dirichlet (GD) distribution defined by [28, 124, 55] has been proposed to smooth this property through two shape parameters and a more general covariance which shows good capabilities for describing proportional data. The GD distribution was introduced as conjugate prior for multinomial distribution to model count data [27]. This distribution arises in various contexts including: Bayesian life-testing problems [55], medical applications such as analysis of acute lymphoblastic leukemia [238], and computer vision like unusual events prediction [239].

In the context of sparse multinomial estimation, there are no recent works even though the interesting utilities in Bayesian statistics, language modeling, graphical models, etc. Motivated by these facts, we propose a novel sparse generalized Dirichlet prior based Bayesian multinomial estimation over large discrete domains. We define prior over exponential hypothesis using vocabulary knowledge; each of which represents a set of feasible outcomes for seen and unseen words. In this work, we consider two different benchmarks to evaluate our proposed approach. A sparse dictionary of German and English poetry where we predict the emotion revealed by each line and the second database concerns tweets allowing us to analyze the flow of emotions related to natural disasters.

7.1.1 Preliminary definitions

With L categories, let $\vec{X}$ be a vector with cell counts $(N_1, \ldots, N_L)$ that follows a multinomial distribution with parameters $(\theta_1, \ldots, \theta_L)$ where N_d represents the number of times in which the d-th category is observed. The probability density function can be represented as follows:

$$p(\vec{X}|\Theta) = \frac{N!}{N_1! \ldots N_L!} \prod_{d=1}^{L} \theta_d^{N_d} \tag{161}$$

where N is the total number of observations.

Assuming a Dirichlet distribution as prior over Θ and using Bayes rule, the posterior density function is given by:

$$p(\Theta|\vec{X}) \propto p(\vec{X}|\Theta)p(\Theta|\vec{\alpha}) \propto \prod_{d=1}^{L} \theta_d^{N_d+\alpha_d-1} \tag{162}$$

Using the Dirichlet prior, the expected $\hat{\theta}_d$ is then given by:

$$\hat{\theta}_d = \frac{\alpha_d + N_d}{\sum_{d=1}^{L} \alpha_d + N} \tag{163}$$

From a predictive Bayesian perspective, the posterior predictive distribution describing beliefs about future observations $\tilde{X}$ is given by:

$$p(\tilde{X}|D, M) = \int_{\Theta} p(\tilde{X}|\Theta, M)p(\Theta|D, M)d\Theta \tag{164}$$

where D is the observed data, M is the model described with the parameters Θ and $p(\Theta|D, M)$ is the posterior distribution for the model parameters.

Given that generalized Dirichlet (GD) distribution is conjugate to the multinomial distribution, a GD prior over Θ is characterized with hyperparameters $(\alpha_1, \ldots, \alpha_{L-1}, \beta_1, \ldots, \beta_{L-1})$ as following:

$$p(\Theta) = \prod_{d=1}^{L-1} \frac{\Gamma(\alpha_d + \beta_d)}{\Gamma(\alpha_d)\Gamma(\beta_d)} \Theta_d^{\alpha_d - 1} (1 - \sum_{j=1}^{d} \Theta_j)^{\delta_d} \tag{165}$$

where $\delta_d = \beta_d - \alpha_{d+1} - \beta_{d+1}, d = 1, \ldots, L-2$ and $\delta_{L-1} = \beta_{L-1} - 1$.

Considering this GD prior, the prediction of initial value of X is :

$$p(X_1 = d) = \frac{\alpha_d}{\alpha_d + \beta_d} \prod_{j=1}^{d-1} \frac{\beta_j}{\alpha_j + \beta_j} \tag{166}$$

Given that if the prior is a GD with hyperparameters $\alpha_1, \ldots, \alpha_{L-1}, \beta_1, \ldots, \beta_{L-1}$, then the posterior is a GD with hyperparameters $\alpha'_1, \ldots, \alpha'_{L-1}, \beta'_1, \ldots, \beta'_{L-1}$ (where $\alpha'_d = \alpha_d + N_d$ and $\beta'_d = \beta_d + \sum_{i=d}^{L} N_i$), the prediction for X_{N+1} is as follows:

$$p(X_{N+1} = d) = \frac{\alpha_d + N_d}{\alpha_d + \beta_d + n_d} \prod_{j=1}^{d-1} \frac{\beta_j + n_{j+1}}{\alpha_j + \beta_j + n_j} \tag{167}$$

where $n_d = \sum_{i=d}^{L} N_i$.

7.1.2 Generalized sparse multinomial

We present in this work a new Bayesian approach for sparse multinomial distribution. Our technique builds up on proposing a new hierarchical prior namely the generalized Dirichlet for the sparse multinomial distribution. We define the new sparse multinomial estimation on two subsets of outcomes: observed and unobserved. The basic strategy of this work can be stated as predicting a new outcome $\vec{X}_{N+1}$ given $D = \{\vec{X}_1, \ldots, \vec{X}_N\}$ drawn from an unknown multinomial distribution.

The Bayesian estimate given a prior over the multinomial distribution can be represented as:

$$p(\vec{X}_{N+1}|D) = \int_{\Theta} p(\vec{X}_{N+1}|\Theta)p(\Theta|\vec{X}_1,\ldots,\vec{X}_N)\, d\Theta \tag{168}$$

where $p(\Theta|\vec{X}_1,\ldots,\vec{X}_N)$ can be obtained using Bayes theorem:

$$p(\Theta|\vec{X}_1,\ldots,\vec{X}_N) \propto p(D|\Theta)\, p(\Theta) \propto p(\Theta) \prod_{d=1}^{L} \theta_d^{N_d} \tag{169}$$

where $p(\Theta)$ is the prior probability of a given Θ.

We note the total vocabulary Σ where a random selection of a subset dictionary $V = \Sigma'$ is consistent with training text data only if it contains all the occurrences d for which $N_d > 0$ where $(\Sigma' \subseteq \Sigma)$. Hence, we refer by Σ^o the set of observed words and $k^o = |\Sigma^o|$.

Given now, a GD prior over possible multinomial distributions for only the observed words with the same hyper-parameters α, β for each word in the vocabulary V ($\forall d \;\; \alpha_d = \alpha, \beta_d = \beta$). Thus, we define the following prior:

$$p(\Theta|V) = \Big[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big]^{|V|} \prod_{d=1}^{L-1} \Theta_d^{\alpha-1}(1-\sum_{j=1}^{d}\Theta_j)^{-\alpha} \tag{170}$$

where $k = |V|,\; k = 1,\ldots,L-1$.

Considering the fact that $\theta_d = 0$ for all $d \notin V$ and using equations 165 and 167, we have:

$$p(X_{N+1} \quad = \quad d|X_1,\ldots,X_N,V) \quad = \quad \begin{cases} \frac{\alpha+N_d}{\alpha+\beta+n_d} \prod_{j=1}^{d-1} \frac{\beta+n_{j+1}}{\alpha+\beta+n_j} & \text{If } d \in V \\ 0 & \text{otherwise} \end{cases} \tag{171}$$

Now, if we use the matter of the uncertainty in vocabularies, this will reflect to expect having words outside the vocabulary. Thus, we assume having a specific probability for the unseen words. First, we define two different priors for all the categories of the set V. Let k be the size of V and assuming a prior that gives equal probability to all the sets with same cardinality.

Given Bayesian multinomial estimation properties, the posterior predictive distribution of a new outcome given the training data D is defined as follows:

$$\begin{aligned} p(X_{N+1} = d|D) &= \sum_V p(X_{N+1} = d|D, V)p(V|D) \qquad (172) \\ &= \sum_{V,|V|=k} p(X_{N+1} = d|D, V)p(V = k|D) \end{aligned}$$

where, using Bayes theorem, $p(V = k|D)$ is given by:

$$\begin{aligned} p(V = k|D) &= \frac{p(D|V = k)p(V = k)}{\sum_{k'} p(D|V = k')p(V = K')} \qquad (173) \\ &\propto p(D|V = k)p(V = k) \\ &\propto \sum_{D,|V|=k} p(D|V)p(V|k)p(V = k) \end{aligned}$$

Next, we assume that we are given the distribution $p(V|k)$ for $k = 1, \ldots, L-1$. So, the prior over sets is:

$$p(V|k) = \binom{L-1}{k}^{-1} \qquad (174)$$

For simplicity purpose, we suppose $p(V = k)$ is the same for all sets of size k that contains Σ^o. Besides, using Bayes rule we can simplify the prediction of X_{N+1}. As there are two cases where any set V has non-zero posterior probability ($d \in \Sigma^o$)

$$\begin{aligned} p(X_{N+1} = d|D) &= \frac{\alpha + N_d}{\alpha + \beta + n_d} \prod_{j=1}^{d-1} \frac{\beta + n_{j+1}}{\alpha + \beta + n_j} \sum_{V,|V|=k} p(V = k|D) \qquad (175) \\ &= \frac{\alpha + N_d}{k^o\alpha + N} \prod_{j=1}^{d} \frac{\beta + n_j}{\alpha + \beta + n_j} \frac{k^o\alpha + N}{\beta + N} \\ &\quad \sum_k p(D|V)p(V|k)p(V = k) \;\; \text{If } d \in \Sigma^o \end{aligned}$$

To simplify the prediction, we move outside the summation terms that doesn't depend on k.

Thus, we need to estimate only:

$$S(D,L) = \frac{k^o\alpha + N}{\beta + N} \sum_k p(D|V)p(V|k)p(V=k) \tag{176}$$

where $S(D,L)$ is considered as the scaling factor and the probability mass assigned to the observed outcomes. Hence, we assign the probability $(1 - S(D,L))$ to unseen words given the training data where $d \notin \Sigma^o$.

Taking advantages of GD properties as a conjugate prior for multinomial distribution for the case of $\Sigma^o \subseteq V$, then we can express the following:

$$\begin{aligned} p(D|V) &= \int_\Theta p(D|\Theta)p(\Theta|V)d\Theta \qquad (177) \\ &= \Big[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big]^{|V|} \prod_{d\in\Sigma^o} \frac{\Gamma(\alpha+N_d)\Gamma(\beta+n_{d+1})}{\Gamma(\alpha+\beta+N_d+n_{d+1})} \end{aligned}$$

Therefore, using the previous equations, we conclude:

$$p(D|V)p(V|k) = \binom{L-1}{k}^{-1} \Big[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big]^{|V|=k} \prod_{d\in\Sigma^o} \frac{\Gamma(\alpha+N_d)\Gamma(\beta+n_{d+1})}{\Gamma(\alpha+\beta+N_d+n_{d+1})} \tag{178}$$

Sampling a random combination to obtain the possible summation over the subsets V, we have

$$\begin{aligned} \sum_k p(D|V)p(V|k) &= \binom{L-1-k^o}{k-k^o}\binom{L-1}{k}^{-1} \frac{\Gamma(\alpha+\beta)^k}{\Gamma(\alpha)^k\Gamma(\beta)^k} \\ &\quad \prod_{d\in\Sigma^o} \frac{\Gamma(\alpha+N_d)\Gamma(\beta+n_{d+1})}{\Gamma(\alpha+\beta+N_d+n_{d+1})} \\ &= \left[\frac{(L-1-k^o)!}{(L-1)!} \prod_{d\in\Sigma^o} \frac{\Gamma(\alpha+N_d)\Gamma(\beta+n_{d+1})}{\Gamma(\alpha+\beta+N_d+n_{d+1})}\right] \\ &\quad \frac{k!}{(k-k^o)!} \frac{\Gamma(\alpha+\beta)^k}{\Gamma(\alpha)^k\Gamma(\beta)^k} \qquad (179) \end{aligned}$$

As the term in the square brackets does not depend on the choice of k, we cancel it out to give rise to the proposed equality:

$$p(V = k|D) = \frac{m_k}{\sum_{k' \geq k^0} m'_k} \tag{180}$$

where

$$m_k = p(V = k) \frac{k!}{(k - k^o)!} \frac{\Gamma(\alpha + \beta)^k}{\Gamma(\alpha)^k \Gamma(\beta)^k} \tag{181}$$

Therefore, the scaling factor is expressed as the following:

$$S(D, L) = \Big(\sum_{k=k^o}^{L-1} \frac{k^o \alpha + N}{\beta + N} m_k \Big) \Big(\sum_{k' \geq k^0} m_{k'} \Big)^{-1} \tag{182}$$

Thus, the simplified prediction probability can be defined as:

$$p(X_{N+1} \quad = \quad d|D) \quad = \quad \begin{cases} \frac{\alpha+N_d}{k^o\alpha+N} \prod_{j=1}^{d} \frac{\beta+n_j}{\alpha+\beta+n_j} S(D, L) & \text{If } d \in \Sigma^o \\ \\ \frac{1}{L-k^o}(1 - S(D, L)) & \text{If } d \notin \Sigma^o \end{cases} \tag{183}$$

7.1.3 Experimental results

Human emotions pose a quite set of challenges that facial expressions are not enough to understand the expressing issues of human being. Concerned about the verbal communication, there is a room for textual information that provides more thoughts about people's opinion and sentiment. So far, limited efforts have focused on sentiment analysis from poetry content for instance poems and books. In this work, we approach the sentiment analysis from various text contents such as poems and messages. Our objectives are to deal with the challenges of text documents that are usually represented with bag-of-words structure (count data) which leads to the sparseness and the high-dimensionality issues. For instance, we consider two challenging applications namely emotion prediction in poetry context and modeling the flow of emotions related to natural disasters.

7.1.4 Emotion prediction in poetry context

In this section, we evaluate the proposed generalized sparse multinomial (GSM) model on public dataset: PO-EMO [1]. The PO-EMO dataset is a collection of German and English language poems which enables modeling aesthetic emotions in poetry. The German corpus contains 158 poems with 731 stanzas written by 51 authors during the period of 1575–1936. The considered poems are extracted from ANTI-K website[1] which provides a platform for student to help them upload edited poems for class including author names, year of publications, poetry topic, and literary epochs. Regarding English dataset, it contains 64 poems (174 stanzas) collected from Project Gutenberg[2] which contains a bunch of eBooks freely available sorted by author, title, topic, language, etc. The two poetry corpora consider emotions elicited in the reader rather than expressed in the text or expected by the author. The emotions are annotated within the context of the whole poem by literary graduate students. Each line is annotated with two labels among the set of considered emotions: Beauty / Joy, Sadness, Uneasiness, Vitality, Awe / Sublime, Suspense, Humor, Annoyance, and Nostalgia (not available in the German data) where the frequencies of these emotions are listed in Table 7.1. We can notice from the Table 7.1 that ”Beauty / Joy” and ”Sadness” are the dominant emotions and the remaining are infrequent which makes this dataset more challenging and interesting to explore the effectiveness of the new sparse multinomial model. It is worthy to mention also that there are no major difference in the emotion frequencies regarding the first annotation and the second one. For that, we consider in our experiments only the first annotation.

The experiments were preformed on Windows 10, an Intel Xeon E-2144G CPU model with a 32 GB RAM and 64-bit operating system. We implemented in Python 3.7 the GSM approach proposed in section 7.1.2, the sparse-multinomials proposed in [237, 235] and [236] for comparison experiments. We consider in our experiments vocabularies with different cardinalities, we give an example of the generated set of words with $L = 300$ for the German and English corpus in Figure 7.1.

In view of evaluating the effectiveness of the proposed generalized sparse multinomial model, we target a challenging scenario in terms of predicting emotions in poetry context. We split the

[1] https://lyrik.antikoerperchen.de/

[2] https://gutentag.sdsu.edu/

Table 7.1: Emotions frequencies for each annotator in the German and English dictionary [1]

	Annotation 1		Annotation 2	
Vocabulary	English	German	English	German
Beauty / Joy	.31	.30	.26	.30
Sadness	.21	.20	.20	.18
Uneasiness	.15	.19	.15	.18
Vitality	.12	.11	.18	.13
Awe / Sublime	.07	.06	.07	.06
Suspense	.04	.07	.07	.08
Humor	.04	.05	.04	.05
Nostalgia	.03	—	.03	—
Annoyance	.03	.04	.02	.02

ein'
es'
ist'
weinen'
und'
das'
der'
welt'
als'
ob'
gott'
aus'
schatten'
liebe'
herzen'
leben'
mauern'
entlang'
eine'
plötzlich'
wie'
aller'
mond'
uns'
wir'
die'
feeling'
better'
blood'
world'
fate'
things'
spring'
wholly'
wisdom'
flowers'
brain'
best'
eyelids'
life'
height'
praise'
arms'
says'
think'
love'
ways'
light'
death'

Figure 7.1: Generated set of words for German and English poems

data into training/testing sets using different ratios (90/10, 80/20, 70/30, 60/40) where each vector $\vec{X}_i$ is a line in the poem corpus. First, we train the model on the training data for the purpose of predicting the poems of the testing set. We represent each poem line as vector of words and we assign a probability for each coming word. If the word is not observed in the context of the previous words, the new word is expected to occur with a specific probability mass. Next, we apply the multinomial mixture clustering on the predicted data to recognize the emotions. We consider the predictive distributions as the multinomial parameters and based on the Bayes rule, we choose the emotion label c corresponding to the poem line $\vec{X}_i$ given by:

$$p(c|D, \vec{\theta}_c) = p(\vec{X}_{N+1}, \ldots, \vec{X}_{N+T}|c)\pi_c \tag{184}$$

where $\{\pi_c\}_{c=1}^{C}$ are the mixing weight coefficients, and $p(\vec{X}_{N+1}, \dots, \vec{X}_{N+T}|c)$ is the c-th probability density function (pdf) of the multinomial distribution that corresponds to the emotion c.

We evaluate the impact of the hyperparameters of the proposed approach using as evaluation metric accuracy and F1-micro scores for comparison analysis. We mention that we assumed that parameters α and β are the same for all the d symbols. Thus, the choice of these parameters will be empirical. For that, we consider three hypothesis: when α and β are equals, when $\beta = 2$, $\alpha \in [0.1, 3]$, and when $\alpha = 2$, $\beta \in [0.1, 3]$. From Figure 7.2, we can deduce the great impact of the parameters α and β on the generalized sparse multinomial and how interfere the relations between the two parameters on the resolution of the model. For the German corpus, we reach the highest accuracy performance when $\alpha < \beta$ but for the English data, the better result is achieved when $\alpha = \beta$. Thus, we conclude that there are no strict relation or value for the parameters and the superior performance is achieved through empirical observations.

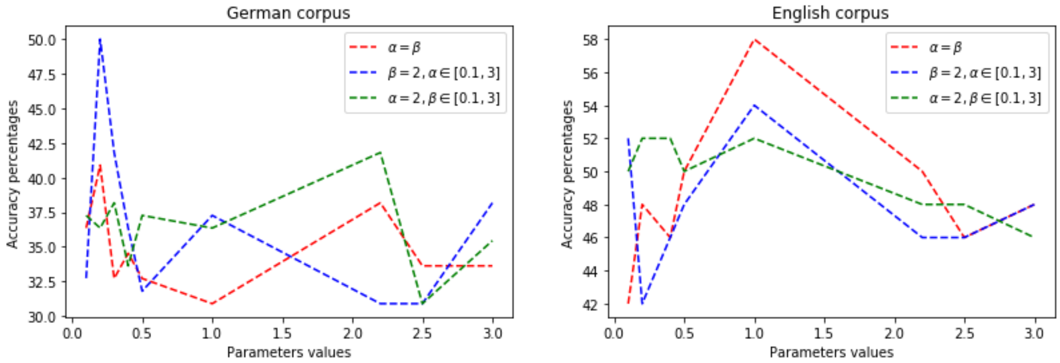


Figure 7.2: Impacts of α and β of the effectiveness of the GSM model

Other important parameters which affect the results are the size of the vocabulary L and the cardinality of the predicted outcomes set T. Besides, we mention that we consider two different priors for the word size: an exponential prior $p(S = k) = \epsilon^k$, and a polynomial prior $p(S = k) = k^{-\epsilon}$, where $\epsilon = 0.9$. Figure 7.3 presents the alteration of the accuracy percentages in terms of L and T. We can see clearly that the performance fluctuates slightly with respect to the size of the vocabulary while it drops quickly with the number of the predicted outcomes. For $L = 200$ and $T = 50$, the GSM achieves the best performance for the two corpora and it should also be underlined that the proposed model is more capable when the prior for the word size is exponential. We mention also that the best results are obtained on a corpus with the total number of words $N = 14,440$. The difficulty of this model occurs when the number of the future outcomes are important. Indeed, this is due to the fact that the prediction of X_{N+t} is based on the preceding prediction $p(X_{N+t-1})$.

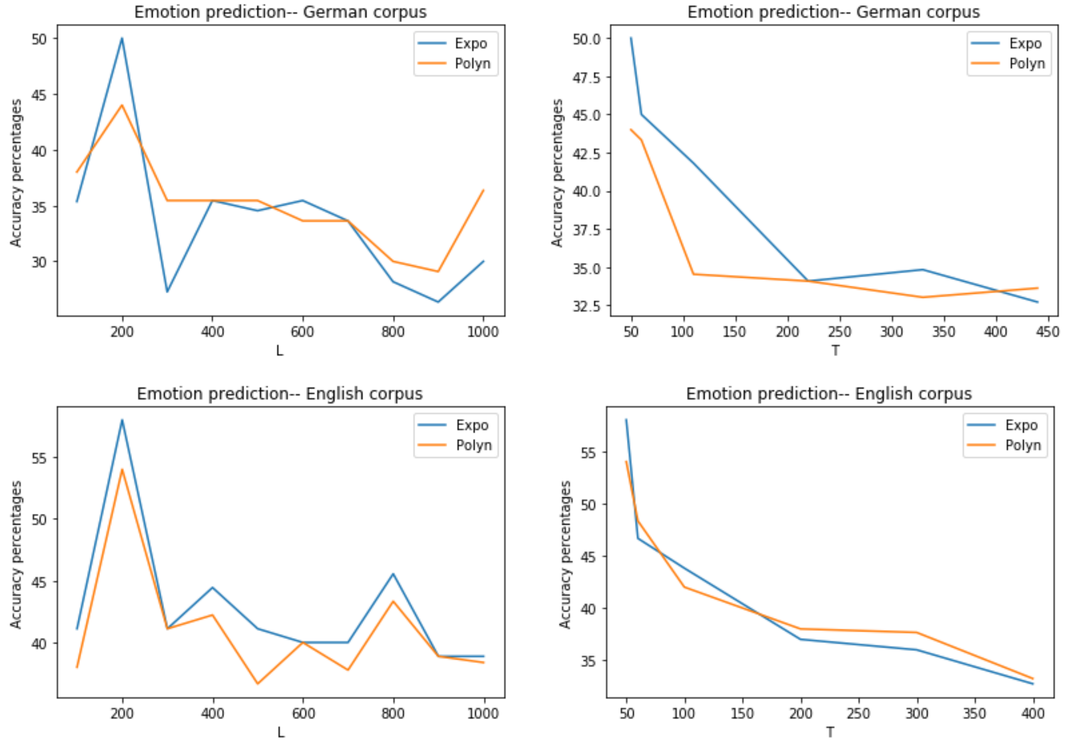


Figure 7.3: Illustration of the emotion prediction results within German and English corpus for different values of L (size of the vocabulary) and T (size of the testing set) using two priors: $P(V = k)$ exponential and polynomial.

The comparison of our proposed model to the baseline methods are summarized in Table 7.2. We refer to the Ristad approach: the model proposed in [235], sparse multinomial: Friedman and Singer's method [237], and the Bayesian sparse multinomial: Griffiths's approach [236]. The entire experiments are averaged over 10 times of running the algorithms. We compare also with BERT model (DBMDZ) applied on the German corpus for line-level in [1]. As illustrated in Table 7.2, the proposed generalized sparse multinomial outperforms all the other related-works on two corpora. The flexibility of the generalized Dirichlet prior gives advantages to the model which results in better performance.

Table 7.2: Comparative results for Line-level emotion prediction

	German		English	
Model	Accuracy	F1-micro	Accuracy	F1-micro
Ristad approach	.33	.28	.36	.25
Sparse multinomial	.33	.31	.37	.27
Bayesian sparse multinomial	.34	.31	.39	.21
BERT-DBMDZ [1]	-	.420	-	-
Generalized sparse multinomial	.50	.425	.58	.419

7.1.5 Modeling the flow of emotions related to natural disasters

We investigate the performance of the proposed Generalized sparse multinomial (GSM) approach on the second application where we consider, CrisisNLP, a crisis corpora collection [145]. The CrisisNLP corpora contains tweets messages related to different natural disasters like earthquake, floods, and infectious disease. In this work, we consider the 2013-Pakistan earthquake database which contains 156,905 tweets messages related to 9 different emotions categories namely: “Injured or dead people”, “Missing, trapped, or found people”, “Displaced people and evacuations”, “Infrastructure and utilities damage”, “Donation needs or offers or volunteering services”, “Caution and advice”, “Sympathy and emotional support”, “Other useful information”, “Not related or irrelevant”. This corpora presents two important challenges as the datasets have high class imbalance and collected messages are introduced as short texts which increase the sparsity problem. For preprocessing, the bag-of-words approach is used to represent the tweets messages as count vectors. We represent each short text “tweets” as an L-dimensional count vector. We display in Figure 7.4 an example of generated vocabulary from the 2013-Pakistan earthquake database with size $L = 600$.

We evaluate the influence of the vocabulary size and the size of testing set on GSM approach and the related works such as Ristard approach (Rt), sparse multinomial (SM), and Bayesian sparse multinomial (BSM). We note from Figure 7.5 the outperformance of GSM with regards to other approaches. We mention that in terms of high-dimensionality (high vocabulary size), the GSM model is able to predict emotions with outsanding performance. It is noteworthy to mention also that the more the vocabulary size is important the more the data is sparse. Thus, the proposed approach succeeds to tackle the challenge of sparsity which proofs the effect of including the vocabulary knowledge for the unseen words. In addition, comparing the GSM with approaches that consider

Figure 7.4: Vocabulary words of 2013-Pakistan earthquake database

the knowledge of unseen words as the SM and the BSM, the GSM achieves the best performance owing to the special characteristics of the generalized Dirichlet prior. Considering the size of testing set, we refer that 200 is 20% of the dataset for testing and 80% for training, 300 represents 30% for testing and 70% for training. We point out that the performance of GSM drops after the size of 600 which is a consequence of the training set is only 30% of the dataset. For that, the performance is affected by the size of the testing-training set of the database.

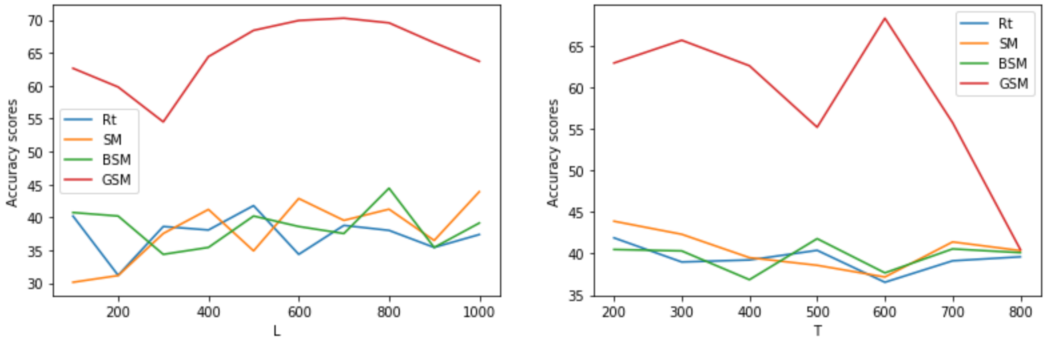


Figure 7.5: Evaluation results of the proposed GSM and comparison with related multinomial-based models in terms of L (vocabulary size) and T (size of testing set)

In Table 7.3, we compare our proposed method in terms of accuracy percentage with the other related multinomial-based models namely: multinomial distribution, Dirichlet compound multinomial (DCM), generalized Dirichlet multinomial (GDM) that are different to Ristard approach (Rt), sparse multinomial (SM), and Bayesian sparse multinomial (BSM). DCM and GDM considers smoothing the multinomial parameters with Dirichlet and generalized Dirichlet priors respectively without Bayesian vocabulary knowledge. These models are not able to deal with this nature of data due to the sparsity and the high-dimensionality issues which demonstrate the effectiveness of including a hierarchical prior with vocabulary knowledge to estimate sparse multinomial.

Table 7.3: Accuracy results on Pakistan earthquake dataset using different multinomial-based models

Models	Accuracy scores
Multinomial	23.81
DCM	28.76
GDM	38.49
Ristard (Rt)	41.79
Sparse Multinomial (SM)	43.91
Bayesian Sparse (BSM)	44.44
Generalized Sparse (GSM)	70.26

7.2 Sparse Document Analysis using Beta-Liouville Naive Bayes with Vocabulary Knowledge

Multinomial Naive Bayes (MNB), mostly used for document classification, is known to be an effective and successful solution [240, 241, 242]. The MNB algorithm is based on Bayes' theorem where the probability of document is generated by multinomial distribution. The strength of this approach lies in the inference speed and the principle of classifying new training data incrementally using prior belief. The MNB was largely applied to sentiment analysis [243], spam detection [244], short text representation [245], and mode detection [246]. However, assuming the features to be independent still presents a limitation for its deployment in real-life applications. Given this central problem in text retrieval, different solutions have been proposed to improve the performance of MNB classifier. Some of these solutions are based on weighting heuristics as in [69], words are associated with term frequencies where common words are given less weight and rare words are accorded to an increased term. Other transformations are based on word normalization such as document length normalization to reduce the influence of common words [247]. Indeed, the most popular heuristic transformation applied to multinomial model is the term-frequency inverse-document-frequency (TF-IDF) associated with log-normalization for all the documents which makes them have the same length and consequently the same influence on the model parameters. Additional alternatives have been applied to enhance the multinomial such as the complement class modeling [69] which solves the problem of unseen words from becoming zero through smoothing the model parameters within a class. Other different smoothing techniques have been associated to MNB such as Laplace smoothing, Jelinek-Mercer (JM), Absolute Discounting (DC), Two-stage (TS) smoothing, and Bayesian smoothing using Dirichlet priors [248, 245, 249].

Among the above approaches, the Dirichlet smoothing gained a lot of attention in information retrieval and different tasks such as novelty detection [250], text categorization [22], texture modeling [26], emotion recognition [11], and cloud computing [25]. With the explosion of online communications, text classification poses huge challenges related to high-dimensional and sparseness problems due to the new variety of available texts including tweets, short messages, customer feedback, and blog spots. Considering large-scale data collections with short and sparse text, text representation based on Multinomial Naive Bayes classifier and the Dirichlet smoothing approaches are not anymore robust enough to deal with these issues. In this context, other priors have been proposed for smoothing the parameters of multinomial such as the generalized Dirichlet distribution [27, 251], the Beta-Liouville [80, 252], and the scaled Dirichlet [103]. By presenting a prior over the training words in a text representation, we ignore the fact that words could be unseen in the vocabulary which correspond to 0 values in the features domain. This fact leads to the sparsity in text categorization. One possible way to cope with this challenge is to exploit a hierarchical prior over all the possible sets of words. In this regards, a hierarchical Dirichlet prior over a subset of feasible outcomes drawn from the multinomial distribution was proposed in [253]. Such knowledge about the uncertainty over the vocabulary of words improves estimates in sparse documents. Yet, still the Dirichlet prior suffers form limitations concerning the strictly negative covariance structure where applying this prior in case of positively correlated data, the modeling will be inappropriate. In this matter, motivated by the potential structure of hierarchical priors [254, 255], we propose a novel hierarchical prior over the multinomial estimates of the Naive Bayes classifier. We introduce the Beta-Liouville prior over all the possible subsets of vocabularies where we consider two assumptions for the “observed” and “unseen” words. We examine the merits of the proposed approach on two challenging applications involving emotion intensity analysis and hate speech detection which are characterized by sparse documents due to the nature of the short text and the large-scale documents.

7.2.1 The Dirichlet smoothing for Multinomial Bayesian Inference

In a Bayesian classification problem, the categorization of an observed data $\mathcal{X}$ with N instances $\vec{X}_1, \dots, \vec{X}_N$ is done by calculating the conditional probability $p(c_j|\mathcal{X})$ for all class c_j. Then, selecting the class that maximizes the following posterior:

$$p(c_j|\mathcal{X}) \propto p(c_j)p(\vec{X}_1, \dots, \vec{X}_N|c_j) \tag{185}$$

where $p(c_j)$ is the proportion of each class c_j and $p(\vec{X}_1, \dots, \vec{X}_N|c_j)$ is the probability that represents the instances within the class. If we consider that the instances are independents within the class c_j, this probability will be given as: $p(\vec{X}_1, \dots, \vec{X}_N|c_j) = \prod_{i=1}^{N} p(\vec{X}_i|c_j)$. Let $\mathcal{X} = (\vec{X}_1, \dots, \vec{X}_N)$ be a set of N independent draws of $\vec{X} = (x_1, \dots, x_{D+1})$ from an unknown multinomial distribution with parameters $(P_1, \dots, P_D)$ defined as follows:

$$p(\vec{X}_i|c_j, \vec{P}_j) = \frac{(\sum_{d=1}^{D+1} x_{id})!}{\prod_{d=1}^{D+1} x_{id}!} \prod_{d=1}^{D+1} P_{dj}^{x_{id}} \tag{186}$$

where $P_{D+1j} = 1 - \sum_{d=1}^{D} P_{dj}$, $\sum_{d=1}^{D} P_{dj} < 1$.

The objective of Bayesian parameter estimation for multinomial distributions is to find an approximation to the parameters $(P_{1j}, \dots, P_{Dj})$ for each class j which can be interpreted as calculating the probability of a possible outcome $\vec{X}_{N+1}$ where:

$$p(\vec{X}_{N+1}|\vec{X}_1, \dots, \vec{X}_N, \zeta) = \int p(\vec{X}_{N+1}|\vec{P}, \zeta)p(\vec{P}|\mathcal{X}, \zeta)d\vec{P} \tag{187}$$

where ζ is the context variables and $p(\vec{P}|\mathcal{X}, \zeta)$ is the posterior probability of $\vec{P}$. By Bayes’ theorem, this conditional probability is given by:

$$p(\vec{P}_j|\mathcal{X},\zeta) = p(\vec{P}_j|\zeta)\prod_{i=1}^{N} p(\vec{X}_i|\vec{P}_j,\zeta) \tag{188}$$

$$\propto p(\vec{P}_j|\zeta)\prod_{d=1}^{D+1} P_{dj}^{M_d} \tag{189}$$

where $M_d = \sum_{i=1}^{N} x_{id}$.

Using the Dirichlet properties, the conjugate prior distribution $p(\vec{P}_j|\zeta)$ is specified by hyperparameters $(\alpha_1,\dots,\alpha_{D+1})$ as follows:

$$p(\vec{P}_j|\vec{\alpha}) = \frac{\Gamma(\sum_{d=1}^{D+1}\alpha_d)}{\prod_{d=1}^{D+1}\Gamma(\alpha_d)}\prod_{d=1}^{D+1} P_{dj}^{\alpha_d-1} \tag{190}$$

Using the fact that when $\vec{P}$ follows a Dirichlet $D(\alpha_1,\dots,\alpha_D;\alpha_{D+1})$ and $p(\vec{X}|\vec{P})$ follows a multinomial distribution then the posterior density $p(\vec{P}|\vec{X})$ is also a Dirichlet distribution with parameters $(\alpha'_1,\dots,\alpha'_D;\alpha'_{D+1})$ where $\alpha'_d = \alpha_d + M_d$ for $d = 1,\dots,D+1$, the estimate of P^*_{dj} is then given by:

$$P^*_{dj} = \frac{\alpha_d + M_{dj}}{\sum_{d=1}^{D+1}\alpha_d + \sum_{d=1}^{D+1} M_{dj}} \tag{191}$$

where M_{dj} is the number of occurrence of x_{id} in document of class j.

7.2.2 The Liouville assumption using vocabulary knowledge

The Liouville family of distributions is an extension of Dirichlet distribution and is proven to be a conjugate prior to the multinomial [256, 257]. The Liouville distribution characterized by positive parameters $(a_1,\dots,a_D)$ and a generating density function $f(.)$ is defined by:

$$p(\vec{P}_j|\vec{a}) = \frac{\Gamma(a^*)}{\prod_{d=1}^{D}\Gamma(a_d)}\frac{\prod_{d=1}^{D} P_{dj}^{a_d-1}}{(\sum_{d=1}^{D} P_{dj})^{a^*-1}} f(\sum_{d=1}^{D} P_{dj}) \tag{192}$$

where $a^* = \sum_{d=1}^{D} a_d$. The probability density function is defined in the simplex $\{(P_{1j},\dots,P_{jD}); \sum_{d=1}^{D} P_{dj} \leq u\}$ if and only if the generating density $f(.)$ is defined in $(0,u)$. A convenient choice for compositional data is the Beta distribution where the resulted distribution is commonly known as Beta-Liouville distribution that presents several well-known properties [79, 78, 80]. To name a few, in contrast to the Dirichlet smoothing, the Beta-Liouville has a general covariance structure which can be positive or negative. Further, with more flexibility the Beta-Liouville distribution is defined with $(a_1, \dots, a_D)$ and two positive parameters α and β:

$$p(\vec{P}_j|\vec{a}, \alpha, \beta) = \frac{\Gamma(a^*)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(\sum_{d=1}^{D} P_{dj})^{\alpha-a^*}(1-\sum_{d=1}^{D} P_{dj})^{\beta-1}\prod_{d=1}^{D}\frac{P_{dj}^{a_d-1}}{\Gamma(a_d)} \tag{193}$$

The Beta-Liouville distribution with $\alpha = \sum_{d=1}^{D} a_d$ and $\beta = a_{D+1}$ is reduced to a Dirichlet $D(a_1, \dots, a_D; a_{D+1})$.

When $\vec{P} = (P_1, \dots, P_D)$ follows $BL(a_1, \dots, a_D, \alpha, \beta)$ and $p(\vec{X}|\vec{P})$ follows a multinomial distribution then the posterior density $p(\vec{P}|\vec{X})$ is a Beta-Liouville distribution with hyperparameters: $a'_d = a_d + M_d$, for $d = 1, \dots, D$, $\alpha' = \alpha + \sum_{d=1}^{D} M_d$, and $\beta' = \beta + M_{D+1}$. According to this property, the Beta-Liouville can be also a conjugate prior to the multinomial distribution where we obtain as a result the estimate of P^*_{dj}:

$$P^*_{dj} = \frac{\alpha + \sum_{d=1}^{D} M_{dj}}{\alpha + \sum_{d=1}^{D} M_{dj} + \beta + M_{D+1}} \frac{a_d + M_{dj}}{\sum_{d=1}^{D}(a_d + M_{dj})} \tag{194}$$

when $\alpha = \sum_{d=1}^{D} a_d$ and $\beta = a_{D+1}$, the estimate according to the Beta-Liouville prior is reduced to the equation 191.

In natural language context, we consider W a corpus of all the used words in $\mathcal{X}$ and the vocabulary $\mathcal{D}$ is a subset of W containing d words. We denote $k = |\mathcal{D}|$ the size of the vocabulary. Using this vocabulary, each text-document is described by an occurrence vector $\vec{X}_i = (x_{i1}, \dots, x_{iD+1})$, where x_{id} denotes the number of times a word d appears in the document i. Considering the text classification problem, we denote W_j is the set of all documents of topic j. For simplicity purposes, we use the same Beta-Liouville parameters for all the words in W_j which gives:

$$P^*_{dj} = \frac{\alpha + M}{\alpha + M + \beta + M_{D+1}} \frac{a + M_{dj}}{ka + M} \tag{195}$$

where $M = \sum_{d=1}^{D} M_{dj}$.

Using vocabulary structure, unseen words are usually presented as zero values which influences the problem of sparsity. In this regard, to give the estimates of multinomial parameters, we use different vocabularies assumptions that solve the problem of unseen words from becoming zero. Assuming W_j^0 is the set of observed words in documents of topic j: $W_j^0 = \{M_{dj} > 0, \forall d\}$ and $k_j^0 = |W_j^0|$. We propose to have two different priors over the values of $\mathcal{D}$. We start with the prior over the sets containing W_j^0 where we assign a Beta-scale that assumes to see all the feasible words under a Beta-Liouville prior and a probability mass function assigned to the unseen words. Thus, by introduction of multinomial estimates, we have:

$$P_{dj}^{*} = p(X_{N+1} = d|W_j, \mathcal{D}) \tag{196}$$

where $\mathcal{D} \subseteq W_j$, allowing to have the following properties:

$$\begin{aligned} p(X_{N+1} = d|W_j) &= \sum_{\mathcal{D}} p(X_{N+1} = d|W_j, \mathcal{D}) p(\mathcal{D}|W_j) \\ &= \sum_{\mathcal{D}, k=|\mathcal{D}|} p(X_{N+1} = d|W_j, \mathcal{D}) p(k|W_j) \end{aligned} \tag{197}$$

and using Bayes'rule, we define the following prior over the sets with size k:

$$p(k|W_j) \quad \propto \quad p(W_j|k)p(k) \tag{198}$$

where:

$$p(W_j|k) = \sum_{W_j^0 \subseteq \mathcal{D}, |\mathcal{D}|=k} p(W_j|\mathcal{D}) p(\mathcal{D}|k) \tag{199}$$

We introduce a hierarchical prior over the sets of $\mathcal{D}$ with size $k = 1, \ldots, D$:

$$p(\mathcal{D}|k) = \binom{D}{k}^{-1} \tag{200}$$

and using the Beta-Liouville prior properties, we define the probability $p(W_j|\mathcal{D})$ as follows:

$$p(W_j|\mathcal{D}) = \frac{\Gamma(ka)}{\Gamma(ka+M)} \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+M)\Gamma(\beta+M_{D+1})}{\Gamma(\alpha+\beta+M')\Gamma(\alpha)\Gamma(\beta)} \prod_{d \in W_j^0} \frac{\Gamma(a+M_{dj})}{\Gamma(a)} \tag{201}$$

Thus summing up over the sets of observed words when $d \in W_j^0$, we have the following probability:

$$\begin{aligned} p(W_j|k) &= \binom{D-k^0}{k-k^0}\binom{D}{k}^{-1} \frac{\Gamma(ka)}{\Gamma(ka+M)} \Gamma(a)^{-k^0} \prod_{d \in W_j^0} \Gamma(a+M_{dj}) \\ &\quad \left[\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+M)\Gamma(\beta+M_{D+1})}{\Gamma(\alpha+\beta+M')\Gamma(\alpha)\Gamma(\beta)}\right] \\ &= \left[\frac{(D-k^0)!}{D!}\Gamma(a)^{-k^0} \prod_{d \in W_j^0} \Gamma(a+M_{dj}) \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+M)\Gamma(\beta+M_{D+1})}{\Gamma(\alpha+\beta+M')\Gamma(\alpha)\Gamma(\beta)}\right] \\ &\quad \frac{k!}{(k-k^0)!} \frac{\Gamma(ka)}{\Gamma(ka+M)} \end{aligned} \tag{202}$$

It is noteworthy to mention that the parameters inside the brackets have no influence on the choice of k. For that, we assume the probability is given by:

$$p(k|W_j) \propto \frac{k!}{(k-k^0)!} \frac{\Gamma(ka)}{\Gamma(ka+M)} p(k) \tag{203}$$

As a result, we have the estimates of Beta-Liouville multinomial parameters for $d \in W_j^0$ defined by:

$$\begin{aligned} p(X_{N+1} = d|W_j) &= \sum_{k=k^0}^{D} \frac{\alpha+M}{\alpha+M+\beta+M_{D+1}} \frac{a+M_{dj}}{ka+M} \\ &\quad \frac{k!}{(k-k^0)!} \frac{\Gamma(ka)}{\Gamma(ka+M)} p(k) \end{aligned} \tag{204}$$

Thus, we define the estimates of Beta-Liouville multinomial parameters using two hypothesis over the “observed” and the “unseen” words as the following:

$$P^*_{dj} = \begin{cases} \frac{\alpha+M}{\alpha+\beta+M'} \frac{a+M_{dj}}{k^0 a+M} \mathcal{B}_l(k,\mathcal{D}) & \text{if } d \in W_j^0 \\ \frac{1}{D-k^0}(1-\mathcal{B}_l(k,\mathcal{D})) & \text{if } d \notin W_j^0 \end{cases} \tag{205}$$

where $M' = M + M_{D+1}$ and $\mathcal{B}_l(k,\mathcal{D})$ is the Beta-scale:

$$\mathcal{B}_l(k,\mathcal{D}) = \sum_{k=k^0}^{D} \frac{k^0 a + M}{ka + M} p(k|W_j) \tag{206}$$

7.2.3 Experimental results

In order to evaluate the proposed approach, we consider two applications that are marked with sparseness of its data namely emotion intensity analysis and hate speech detection; both of them in tweets. We compare the novel Beta-Liouville Naive Bayes with vocabulary knowledge so-called (BLNB-VK) with the related-works: Multinomial Naive Bayes (MNB), the Dirichlet smoothing (DS), the Dirichlet smoothing with vocabulary knowledge (DS-VK), and the other models applied on the considered datasets. In our experiments, we assign an exponential prior for the word size $p(k) \propto \gamma^k$, where we set $\gamma = 0.9$. The results obtained in these two applications are achieved with the following hyperparameters: $a = 1$, $\alpha = 0.2$, $\beta = 0.2$. Details on the implementation code are available on .

Emotion intensity analysis

With the explosion of the new communications means, speaking out our feeling takes new formats than verbal and facial expressions. Social media allows people to share their emotions, opinions, and even their attitudes. These online expressions give the opportunity for entrepreneur, producer, and politician to predict the preference and the tendency of their community. In our work, we consider a dataset that considers not only emotions in tweets but also the intensity dataset [258]. The tweet emotions intensity (EmoInt) dataset contains four focus emotions: *anger, joy, fear*, and *sadness* where tweets are associated with different intensities of each emotion. For example, for the subset *anger*, 50 to 100 terms are associated with that focus emotion as *mad, frustrated, furry,* *peeved*, etc. The EmoInt dataset [3] contains $7,097$ of tweets split into training, development, and testing sets. We consider in our experiments the testing set for comparison issues and we randomly split the data into $70/30$ to generate the vocabulary. In Figure 7.6, we show a vocabulary of $352,000$ words with 600 features size.

Figure 7.6: Visualization of the vocabulary of words for EmoInt dataset

We explore the influence of the features dimensions and the number of words on the performance of the proposed approach. Accordingly, we measure the accuracy of classifiers. Figure 7.7 shows the outperformance of the BLNB-VK when the feature of texts are less than 200 and when we have also documents with more than 1000 size of vocabulary. This proves the ability of our algorithm to recognize short texts as well as high-dimensional documents. We mention that number of words affect also the classification of the tweets emotions where we have a vocabulary with less words is not able to recognize properly the emotions intensity. Yes, with an adequate number of words as $150,000$, the tweets are accurately detected by the mentioned classifiers: BLNB-VK, DS-VK, DS, and MNB. In Table 7.4, the performance of different classifiers is compared according to Pearson correlation and accuracy percentage where the best overall results are again achieved by BLNB-VK.

[3] http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html

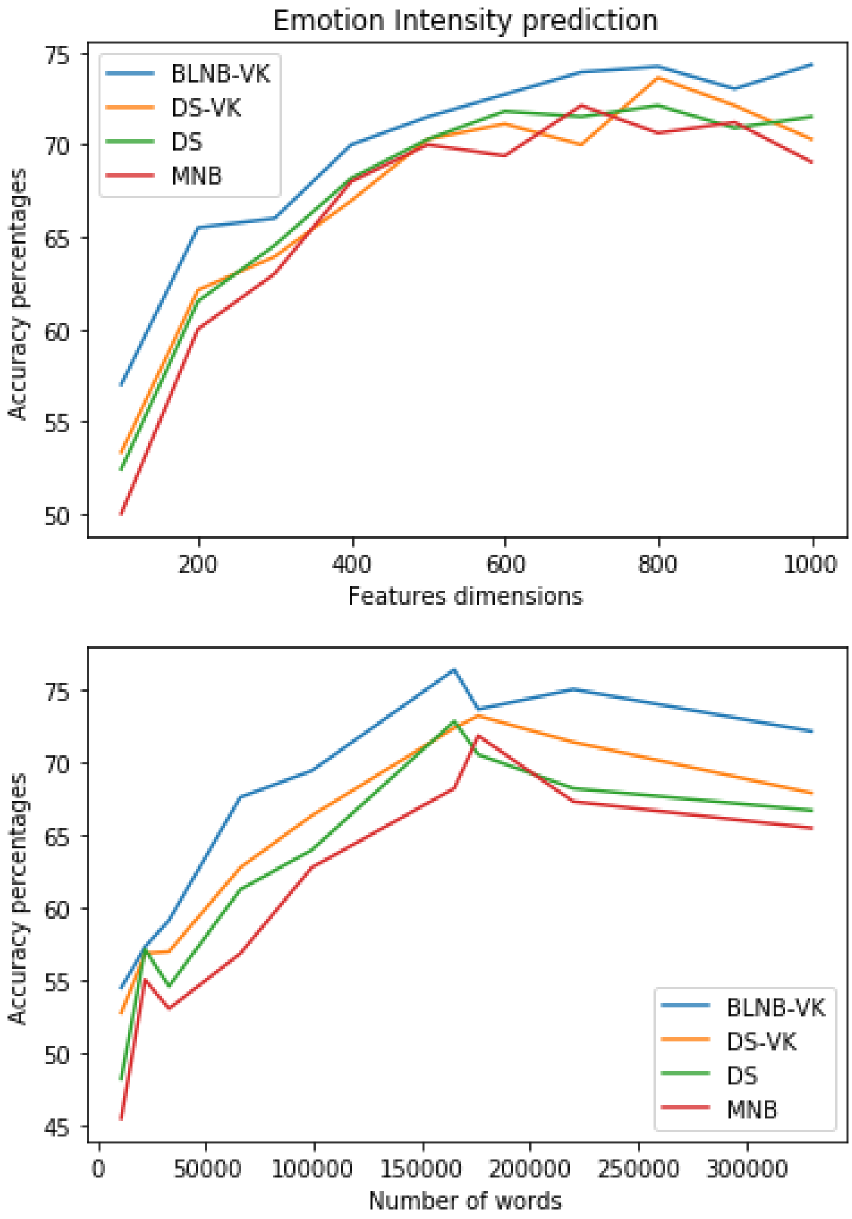


Figure 7.7: Influence of feature dimension and the number of words on the performance of the proposed approach BLNB-VK and comparison with Naive-based related works on EmoInt dataset

Hate speech detection

Recently, social media becomes a platform of expressing racism, religious beliefs, sexual orientation, and violence. Violence and related crimes are on rise due to the spreading of online hate speech. Researchers and social media companies as Facebook and Tweeter conduct efforts to detect

Table 7.4: Pearson (Pr) correlations and Accuracy percentages of emotion intensity predictions using different classifier methods

Model	Pr correlation	Accuracy
Word embedding (WE) [258]	0.55	-
Lexicons (L) [258]	0.63	-
WE + L [258]	0.66	-
MNB	0.59	72.12
DS	0.61	71.51
DS-VK	0.64	73.63
BLNB-VK	0.68	76.36

and delete offensive materials. We are interested in this type of data where haters express their beliefs in short tweets or messages. In this regard, we study the performance of our proposed approach on Tweeter hate speech dataset [259]. The dataset contains $24,802$ tweets categorized into hate speech, offensive but non-hate, and neither. Each tweet was encoded by three experts based on Hate-base lexicon where only 5% were encoded as hate speech, 76% are considered to be offensive, and the remainder are considered neither offensive nor hate-speech. We split the dataset into training and testing sets to form the vocabulary.

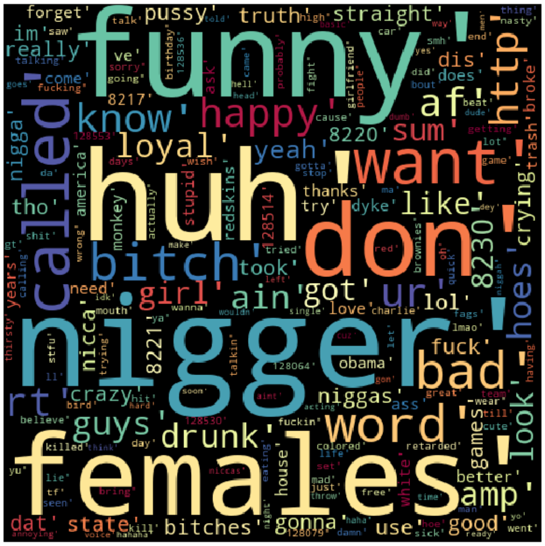


Figure 7.8: Visualization of the vocabulary of words for Hate speech dataset

Figure 7.8 shows an example of set of words observed in the vocabulary of the dataset. By comparing performance results in Figure 7.9, we can see that size of the vocabulary and the overall number of words have influenced the classification of hate speech texts. Indeed, we mention that

BLNB-VK achieves the highest classification accuracy while Multinomial Naive Bayes attains the least performance in both cases: short texts and high-dimensional documents. We compare also the classification results against previously reported results on the same dataset (Table 7.5 where our proposed approach outperforms the Logistic regression in [259] by 10% with respect to accuracy metric.

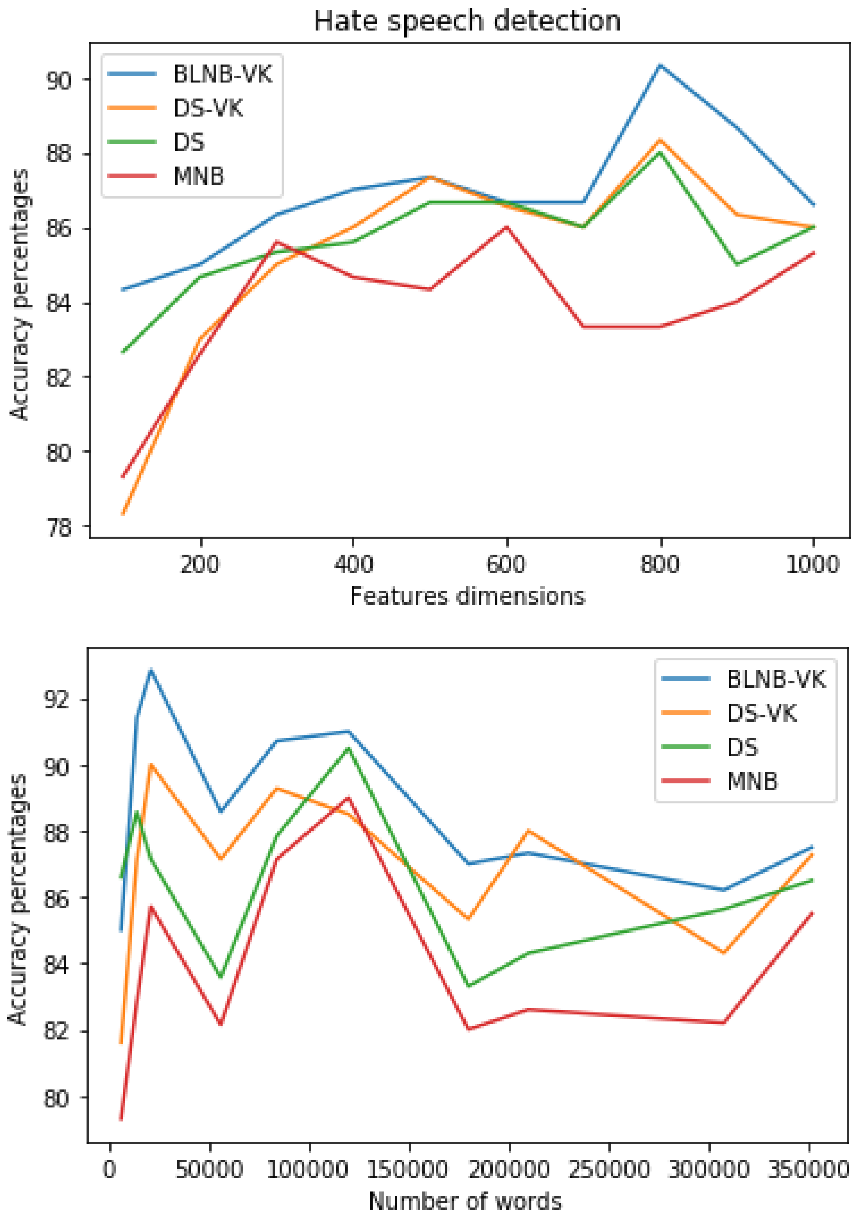


Figure 7.9: Influence of feature dimension and the number of words on the performance of the proposed approach BLNB-VK and comparison with Naive-based related works on hate speech dataset

Table 7.5: Classification results of different methods on hate speech dataset

Model	Accuracy
Logistic regression [259]	82.33
MNB	89
DS	90.5
DS-VK	90
BLNB-VK	92.85

7.3 Conclusion

In this chapter, we presented a hierarchical prior for the problem of estimating the sparse multinomial parameters. We proposed a novel sparse model based on the generalized Dirichlet prior which reduces the limitations of Bayesian approaches previously introduced for language model. Our method exploits the knowledge of the structure of the vocabularies in natural language. Explicitly, we employ prior knowledge for two feasible sets of vocabulary of words to be used for the estimation of sparse multinomial distribution. We consider predicting each new word if it is observed/not observed with a specific probability mass. We showed the effectiveness of the proposed approach in predicting emotions in German and English poetry and modeling the flow of emotions related to natural disasters. Different parameters influence the performance of the GSM algorithm, demonstrating that it can be properly considered to more complex applications. Although the proposed model achieved interesting and competitive results, we aim for a future work to investigate its limitations with respect to the size of testing set.

Additionally, we investigated the problem of sparse document analysis using Bayesian classifier. First, we presented an alternative Liouville prior for the Multinomial Naive Bayes: the Beta-Liouville distribution. Then, we incorporated the vocabulary knowledge to BLNB for the purpose of taking into account the unseen words in the vocabulary. We introduced hierarchical priors over all the possible sets of the vocabulary. Evaluation results illustrate how our proposed approach was able to analyze emotion intensities and to detect hate speech in Tweeter. Better results have been obtained by the proposed BLNB-VK with comparison to the other related Bayes classifier, for instance, Multinomial Naive Bayes, Dirichlet smoothing, and Dirichlet smoothing with vocabulary knowledge in both applications. In our approach, we assume that all parameters are the same for all words for inference simplicity. However, fixing model's parameters could affect the flexibility. For that, a promising extension of this paper could be to investigate the inference of parameters of the BLNB. Further, our proposed model could be combined with reinforcement learning where the probabilities of each state could use Beta-Liouville multinomial parameters.

Chapter 8

Interactive Distance Dependent IBP Compound Dirichlet Process

The sparse topic model (sparseTM) is a nonparametric Bayesian model that employs a Spike and Slab prior which decouples sparsity and smoothness in the topic mixtures. Though powerful, sparseTM considers the sparsity problem only for the topic-word distributions where sparsity-enhanced topic models emerged to extract focused topics and focused terms. However, these models assume that data are exchangeable which often fails for real-world data where dependencies between features are expected. We present a generalization of the distance dependent IBP, the interactive distance dependent Indian buffet process compound Dirichlet process (idd-ICDP), for modeling non-exchangeable text data. To the best of our knowledge, idd-ICDP is the first nonparametric Bayesian model supporting non-exchangeable data applied to sparse topic models. The idd-ICDP allows an interactive framework integrating human experts knowledge with potentially an unbounded number of topics and vocabulary words in a corpus. We derive a Markov Chain Monte Carlo sampler combined with Metropolis-Hastings algorithm and study its performance on benchmark corpora and sentiment analysis data. Experiments demonstrate that accounting the non-exchangeablility nature of real-world data gives better predictive performance and that the interactive strategy offers better high-quality topics.

8.1 Introduction

With the rapid spread of online information in different fields, probabilistic topic models such as probabilistic Latent Semantic Analysis (PLSA) [260] and Latent Dirichlet Allocation (LDA) [181] have been developed to explore and analyze large text corpora [261, 262, 263]. These classical topic models generally regard a collection of documents as a mixture of a fixed number of topics where each topic is a discrete distribution over words and each word is assumed to be generated from one of these topics. Compared to PLSA, LDA introduces a smoothing process over the document-topic distributions and the topic-word distributions drawn from a Dirichlet prior.

Despite the success of LDA for topic modeling, it is noteworthy to mention that it adopts the bag-of-words representation of the documents and it ignores the sequential structure of text. As a result, classical models fail to overcome the sparsity problem in real-world applications [264, 265]. Instead of defining topic mixtures in the entire simplex, Wang et al. [266] handle the sparsity of topic mixtures along with a smoothing prior and propose the sparse topic model (sparseTM). They apply a "Spike and Slab" prior that decouples smoothness and sparsity in which topics are defined on a random subset of the vocabulary (addressing sparsity) and then only the selected terms are smoothed. Allowing the smoothness of topic simplex, a focused topic model (FTM) [267] was proposed as a new sparse topic model to handle the sparsity in document-topic distributions. In the FTM, focused topics reflect the fact that topic-document distribution is focused on a sparse subset of topics with a specific document. Mining focused topics and focused terms, a dual sparse topic model was introduced for short text [268]. The dual sparse topic model (DsparseTM) addresses the sparsity in both the document-topic mixtures and topic-word distributions. DsparseTM introduces focused topics and focused terms to restrict the size of the topic simplex and the word simplex. Despite addressing the sparsity in both topic mixtures and word usage, DsparseTM considers a fixed number of topics. Accordingly, a major limitation of probabilistic topic models is considering fixed number of topics. Reconsidering the bounded topic assumption, the Hierarchical Dirichlet process (HDP) [269] is often used in Bayesian nonparametric topic modeling [270, 271]. For instance, the sparseTM was built on the HDP where the number of topics is unbounded. Another well known nonparametric Bayesian model is the Indian Buffet process (IBP) typically used for latent feature models [272]. The IBP was inspired by the Indian buffet where customers can choose an infinite number of dishes. The FTM combines HDP with infinite binary matrices constructed using IBP. An extension of IBP that combines the properties of HDP and the IBP was proposed in [273], the IBP compound Dirichlet Process (ICD) applied to focused topics. However, FTM was based on the one-parameter IBP which is not power-low distributed and generate sparse binary matrices. Accordingly, Archambeau et al. [2] introduced a four-parameter IBP compound Dirichlet process (ICDP) to sparse nonparametric topic modeling called latent IBP compound Dirichlet allocation (LIDA). ICDP allows more flexibility and exhibits power-law characteristics for topic modeling.

However, all the above mentioned sparse models do not consider dependencies between data points and assume that data are exchangeable. Broadly, one more shortcoming of traditional topic models is assuming that documents are exchangeable: permuting the order in which they appear leaves the probability of seeing a topic unchanged. Though, this assumption is unreasonable for textual data where we expect topics to be auto-correlated over time or more generally covariate-dependent. To relax this assumption, variants of Dirichlet process adapt nonparametric models to non-exchangeable data that allow dependencies between documents [274], [275], [276], [277]. Distance dependent Chinese restaurant process (dd-CRP) [278] introduces dependencies between data elements over time, space, and network connectivity in infinite models. dd-CRP connects customers to others customers while the traditional CRP connects customers to tables. The distance dependent CRP allows accommodating the non-exchangeability of the data in which customers assignment depends on the distances between them that could be over time, space, or other characteristics. A generalization to the IBP and the dd-CRP was proposed in [279], the distance dependent IBP (dd-IBP). The dd-IBP extends dd-CRP to infinite latent feature models and allows as well to capture non-exchangeable structure.

On social media platforms, user-generated content (UGC) is reshaping e-commerce and introducing a new marketing strategy. UGC offers the opportunity for organisations in different field to understand customer behaviors. Such content can mostly be presented as unstructured text where topic models have been used to extract hidden semantic information. Although LDA and its variants provide a powerful tool for modeling such text data, the discovered topics are not always interpretable by users and some of the words does not make sense for customers which affect topic quality. Recently, an interactive latent Dirichlet allocation (iLDA) was proposed [280] to obtain subjective topic-word distribution from human experts. Integrating human knowledge can draw high-quality topics for real-word corpora. The proposed iLDA presents a useful model not only for UGC but also for all kind of text data to improve topic quality.

In this work, we introduce an interactive distance dependent IBP compound Dirichlet process (idd-ICDP) to address the following issues:

(1) The sparsity problem in textual data, in particular, count data. We propose a Spike and Slab prior in our model where we smooth topic-word and topic-document distributions by introducing Bernoulli variables over words and topics, respectively.

(2) The bounded topics and words assumption. We introduce a novel nonparametric Bayesian model and applied, for the first time, to topic modeling. We adapt the dd-IBP to topic modeling and propose the new dd-IBP compound Dirichlet process (dd-ICDP).

(3) The dependency between features over time and the exchangeability data assumption. We introduce the dd-ICDP for the purpose of considering the non-exchangeability of data through sampling topics/words assignments using the distance between them over time.

(4) Integrating human experts knowledge for the purpose of improving topic quality. We present an interactive dd-ICDP (idd-ICDP) using an objective topic-word distribution generated from dd-ICDP and a subjective topic-word distribution based on human experts' belief.

The remainder of this chapter can be summarized as follows. In Section 2, we introduce sparsity and smoothness in topic models and we present the generative process of SparseTMs briefly. Next, in Section 3, we present the distance dependent IBP for latent feature models and we explain a small example of a latent feature matrix generated by dd-IBP. Section 4 provides an introduction to the interactive LDA. We then develop the novel interactive nonparametric Bayesian model so-called interactive distance dependent IBP compound Dirichlet process where we present the process of topic generation and document generation, and the complete generative process in Section 5. We explore the inference of each parameter in Section 6 using MCMC sampling approach. In Section 7, we validate our model, compare its performance to the state-of-arts models on several benchmark

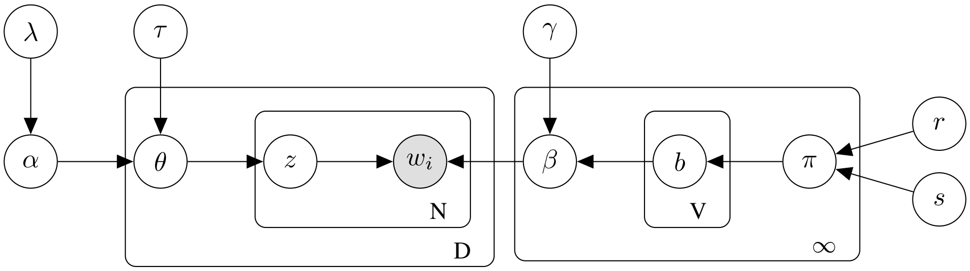


Figure 8.1: The graphical model of sparseTM.

corpora and consider the proposed model as a supervised topic mooel for sentiment analysis. We conclude this paper in Section 8 with discussions and some future insights.

8.2 Sparsity and Smoothness in Topic models

The key idea of decoupling sparsity and smoothness in topic models is to introduce Bernoulli variables over words that represent “on” if the term appears in the topic and “off” otherwise. This fact was the main goal of Sparse topic models (sparseTMs) which is known also as “Spike and Slab” prior that has been used in many real world applications [281, 282, 283, 284, 285]. SparseTMs permit controlling in an independent manner the number of terms in topics to address the sparsity problem and regulating the probability of words which is responsible of the smoothness issue. Decoupling sparsity and smoothness in topic modeling was first proposed in [266] using a Dirichlet distribution over the topics and Bernoulli variables over words named “term selector” to indicate whether a term is selected by the topic. Extending the work to Dual-Sparse topic model for short texts, the authors in [268] address the sparsity in both topics and terms by introducing “focused topics” and “focused terms”. The graphical model representation for sparseTM is depicted in Figure 8.1 where b is the term selector, π is the term selection proportion and the topic distribution is drawn over Dirichlet distributions as follows:

$$\begin{aligned} \beta &\sim \text{Dirichlet}(\gamma \mathbf{b}), \\ b_{kv} &\sim \text{Bernoulli}(\pi_k), \\ \pi_k &\sim \text{Beta}(r, s), \end{aligned}$$

The strength of sparseTM is controlling the sparsity of the topic with a proportion of zeros in a bank of Bernoulli random variables and enforcing the smoothing over the terms with non-zeros b_{kv} through the hyperparameter γ. The sparsity of a topic k is defined as follows:

$$\text{sparsity}_k \quad = \quad 1 - \sum_{v=1}^{V} b_{kv}/V. \tag{207}$$

8.3 Distance dependent IBP for latent feature models

The Indian Buffet process (IBP) is a nonparametric Bayesian model introduced to generate unbounded latent features when data are assumed to be exchangeable. The dd-IBP [279] is a prior over binary latent feature matrices **Z** with an infinite number of columns and a finite number of rows. Like the IBP, rows of **Z** correspond to customers and columns correspond to dishes. In dd-IBP, a distance in time or space is associated for each pair of customers. The generative process of features is defined as a sequential process. The first customer enters the restaurant and selects a Poisson-distributed number of dishes, where the dishes selected by a customer are defined as “owned” by this customer. A dish can be owned by one customer or unowned. Next, for each owned dish, unlike IBP where customers sample new dishes, customers connect to other customers. Thus, a graph of connections is defined between customers where dishes are inherited by a customer if there exist a path to reach the dish’s owner. The active features in this model are those that each customer owns or inherits.

We display an example of per-dish graph for customers assignments in Figure 8.2. In this example, dish 1 is owned by customer 1; customer 3 reaches customer 1 for dish 1 directly and customer 4 inherits the dish 1 through a chain. Thus, feature 1 is activated for customers 1, 3, and 4. Customer 2 owns dish 2 and customer 1 reaches customer 2 for dish 2. Dish 3 is owned by customer 3, customers 1, 2, 4 reach customer 3 for dish 3, either directly or through a chain. Hence, dish 3 is activated for all the customers. Dish 4 is owned by customer 1 and only customer 3 reaches customer 1 for dish 4.

To generate the latent feature matrix **Z**, a connectivity matrix $C = (c_{ij})$ is defined which associates each dish with a set of customer-to-customer assignments where $c_k^* = i$ indicates that

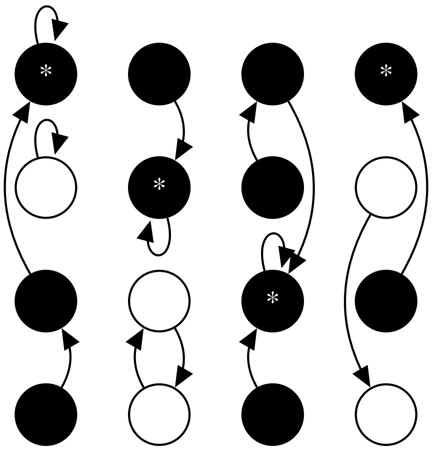


Figure 8.2: Example of a latent feature matrix generated by dd-IBP. Rows correspond to customers and columns correspond to dishes. Black shading indicates that a feature is active for a given customer.

customer i owns dish k. Next, in order to sample dishes, the distance matrix between customers $D = (d_{ij})$ plays an important role where the customer assignment is determined according to the probability that customer i connects to customer j for dish k given a *decay function*: $p(c_{ik} = j|D, f) = a_{ij}$. Hence, we obtain a normalized proximity matrix $A = (a_{ij})$. The *decay function* f requires that $f(0) = 1$ and $f(\infty) = 0$ which controls the probabilities of customers that we call proximity: $a_{ij} = f(d_{ij})/h_i$, where $h_i = \sum_{j=1}^{N} f(d_{ij})$. The generative process of the dd-IBP defines the joint distribution of the ownership vector as follows:

$$p(C, c^*|D, \alpha, f) = p(c^*|\alpha)p(C|c^*, D, f). \tag{208}$$

where the first term is the probability of the ownership vector:

$$p(c^*|\alpha) = \prod_{i=1}^{N} p(\lambda_i|\alpha), \tag{209}$$

where $\lambda_i \sim$ Poisson(α/h_i) is the Poisson distributed number of dishes for each customer i and the second term is defined from the conditional distribution of customer assignments:

$$p(C|c^*, D, f) = \prod_{i=1}^{N} \prod_{k=1}^{K} a_{ic_{ik}}. \tag{210}$$

8.4 Interactive LDA

The iLDA model [280] combines the knowledge of human expertise to generate subjective topic-word distribution with the objective topic-word distribution mined by LDA. We suppose we have D documents where each document d consists of a sequence of words $\vec{w}_d = \{w_{d1}, \dots, w_{dV}\}$ defined in a vocabulary of size V and let n_d denote the nth word in document d. Suppose the subjective topic-word distribution generated by human knowledge is ϕ_u and the objective distribution ϕ_l, iLDA defines the topic-word distribution as a linear weighted sum: $\phi = \lambda_1 \phi_l + \lambda_2 \phi_u$. In this section, we describe the calculation of the subjective topic-word distribution for both deterministic and stochastic strategy.

For deterministic strategy, the objective topic-word distribution is based on the degenerative probabilities given by experts and considers always human experts' knowledge accurate and reliable. Thus, the adjusted topic-word distribution is calculated as:

$$\phi_{uT'}^{(k)} = \{p_{lt'}^{(k)} \times p_{ut'}^{(k)} | t \in W_T, t' \in W_{T'}\} \tag{211}$$

where W_T represents the T most probable words in topic k, $W_{T'}$ represents the T' selected words by human experts with probabilities $p_{lT'}^{(k)}$, and $p_{ut'}^{(k)}$ is the degenerative probability that measures the distrust of human experts for word t' in topic k. In order to validate that the sum of the probabilities of all words in a topic are equals to 1, the other most probable words in $(W_T - W_{T'})$ are adapted by the surplus probabilities from the adjusted words as:

$$\begin{aligned} \phi_{u,-T'}^{(k)} \quad = \quad & \Bigg\{ p_{-t'} \times \left(1 + p_r \frac{p_{-t'}}{\sum_{-T'} p_{-t'}}\right) | - t' \in W_T \\ & \text{and } - t' \neq W_{T'} \Bigg\}, \end{aligned} \tag{212}$$

where $p_{-t'}$ are the probabilities in $W_{T'}$ and $p_r = \sum_{t'=1}^{T'} p_{t'} \times (1 - p_{ut'}^{(k)})$ measures the surplus probability. Hence, the subjective topic-word distribution obtained by the deterministic strategy is $\phi_u^{(k)} = \{\phi_{uT'}^{(k)}, \phi_{u,-T'}^{(k)}, \phi_{V-T}^{(k)}\}$.

The stochastic strategy adjusts the topic-word probabilities for the words while taking into account human experts' knowledge. The degenerative probabilities are controlled with a signal function where a higher probability means human experts largely accept the objective topic-word distribution while a smaller value reflects accepting the adjusted probabilities according to the belief of human experts. Accordingly, a stochastic variable u is defined to determine if the probability words in a topic should be adjusted by human experts:

$$\phi_{uT'}^{(k)} = \{p_t^{(k)} \times [p_{t'}^{(k)}]^{I(u>p_{t'}^{(k)})} | t \in T, t' \in T'\} \tag{213}$$

where $I(u > p_{t'}^{(k)})$ is the signal function which is equals to 1 if $u > p_{t'}^{(k)}$ and 0 otherwise. The degenerative probabilities of the words are weighted as in the deterministic strategy (equation 212) and hence the subjective topic-word distribution is obtained.

8.5 Interactive Distance Dependent IBP Compound Dirichlet Process

The proposed idd-ICDP focuses on accommodating the sparsity issue through the "Spike and Slab" prior over Dirichlet distributions for the topics and terms. Our model can be seen as infinite spike and slab model where we assume that the number of topics and the number of vocabulary words are unbounded using a Bayesian nonparametric prior on the topic and the word proportion matrices. To capture the non-exchangeability structure of the data and characterize its feature-sharing properties, we propose the dd-IBP compound Dirichlet process (dd-ICD), an alternative Bayesian nonparametric prior to the Dirichlet process that combines properties from the HDP and dd-IBP. We apply the dd-ICDP as a prior for document-topic distribution and topic-word matrix. We introduce also, in this work, an interactive strategy to mine high-quality topics through integrating human experts knowledge. The notations that we use in our model are summarized in Table 8.1.

The distance dependent Indian Buffet process for documents and topics generation can be summarised as follows:

Table 8.1: Variables and Notations

Notation	Meaning
K	number of topics
V	vocabulary
D	collection of documents
N_d	length of document d
B	connectivity matrix for topics
M	distance matrix between topics
A	normalized proximity matrix for topics
f	decay function
$\mathcal{R}$	topics reachability
C	connectivity matrix for documents
Q	distance matrix between documents
G	normalized proximity matrix for documents
$\mathcal{L}$	documents reachability
β_k	term selector
τ	DP hyperparemeter
Φ	topic distribution over words in interactive model
Φ^s	topic-word distribution generated by expert knowledge
Φ_k^u	topic-word distribution generated by the model
$\bar{\Phi}_k^u$	word activation per word
λ_1	the trust of the model
λ_2	the trust of human beings
θ_d	document-topic distribution
α	DP hyperparameter
η	dd-IBP parameter
γ	deterministic many-to-one function
Θ	document distribution over topics
$\bar{\theta}_d$	topic activation per document
π_k	topic selector
η	dd-IBP parameter
w_{di}	i-th word in document d
z_{di}	topic of the i-th word in document d

8.5.1 Topic generation

a. **Assign words ownership:** Each topic k selects a Poisson-distributed number of words v, let $\sigma_k \sim \text{Poisson}(\delta/y_k)$, thus allocating σ_k words to this topic and set the ownership $b_v^* = k$. The total number of owned words is $\mathbb{K} = \sum_{k=1}^{K} \sigma_k \sim \text{Poisson}(\delta/y)$, where $y = \sum_{k=1}^{K} y_k$

b. **Assign topic connections:** Each word is associated with a set of topic-to-topic assignments, specified by $V \times \mathbb{K}$ *connectivity matrix* **B**, where $b_{vk} = k'$ indicates that topic k connects to topic k' for word v. Given **B**, topics form a set of directed graphs according to $p(b_{vk} = k'|M, f) = a_{kk'}$, where $\mathbf{M} = (m_{kk'})$ is the *distance matrix* between topics and f is the decay function. $f : \mathbb{R} \mapsto [0, 1]$ maps distance to a quantity which controls the probabilities of topics. By applying the decay function to each topic and normalized by topics, we define the *normalized proximity matrix* for topics **A**, $a_{kk'} = f(m_{kk'})/y_k$ where $y_k = \sum_{k'=1}^{K} f(m_{kk'})$.

c. **Compute words inheritance:** A topic k inherits a word v if there exists a path along the directed graph for the word v from the topic k to the word's owner b^*_v. The owner of a word automatically inherits it. We encode *reachability* with $\mathcal{R}$. If topic k is reachable from k' for word v, $\mathcal{R}_{kk'v} = 1$, otherwise $\mathcal{R}_{kk'v} = 0$.

d. **Compute topic-word indicator matrix:** For each topic k and word v, we set $\Phi_{kv} = 1$ if k inherits v, otherwise $\Phi_{kv} = 0$.

8.5.2 Document generation

a. **Assign topics ownership:** Each document d selects a Poisson-distributed number of topics k, let $\mu_d \sim$ Poisson(η/h_d), thus allocating μ_d topics to this document and set the ownership $c^*_k = d$. We define the total number of owned topics by $\mathbb{D} = \sum_{d=1}^{D} \mu_d$.

b. **Assign document connections:** Each topic is associated with a set of document-to-document assignments, specified by $K \times \mathbb{D}$ *connectivity matrix* **C**, where $c_{kd} = d'$ indicates that document d connects to document d' for topic k. Given **C**, documents form a set of directed graphs according to $p(c_{kd} = d'|Q, f) = q_{dd'}$, where $\mathbf{Q} = (q_{dd'})$ is the *distance matrix* between documents and f is the decay function defined as in the topic generation. By applying the decay function to each document and normalized by documents, we define the *normalized proximity matrix* for documents **G**, $g_{dd'} = f(q_{dd'})/h_d$ where $h_d = \sum_{d'=1}^{D} f(q_{dd'})$.

c. **Compute topics inheritance:** A document d inherits a topic k if there exists a path along the directed graph for the topic k from the document d to the word's owner c^*_k. The owner of a topic automatically inherits it. We encode *reachability* with $\mathcal{L}$. If document d is reachable from d' for topic k, $\mathcal{L}_{dd'k} = 1$, otherwise $\mathcal{L}_{dd'k} = 0$.

d. **Compute document-topic indicator matrix:** For each document d and topic k, we set $\Theta_{kd} = 1$ if d inherits k, otherwise $\Theta_{kd} = 0$.

8.5.3 Generative process

We suppose that we have a collection of documents $D = \{\vec{w}_d\}_{d=1}^{|D|}$, where each document d is a sequence of words representing its textual content $\vec{w}_d = \{w_{d1}, \dots, w_{dN_d}\}$, where w_{di} denotes the frequency of the i-th term in document d, and N_d is the size of document d and the total number of words in the corpus is given by $N = \sum_d N_d$. The graphical representation of the proposed model is depicted in Figure 8.3. The formal definition of the generative process of idd-ICDP is as follows: For each topic $k \in \{1, 2, \dots\}$:

- For each term $v \in \{1, 2, \dots\}$;
  - ○ Sample word activator $\bar{\phi}_{kv}^{u} \sim$ dd-IBP(δ);
  - ○ Sample word distribution:
    $\Phi^u \sim$ dd-ICDP(δ, τ), $\Phi^u | \bar{\Phi}^u \sim$ DP($\tau \bar{\Phi}$);

For each document $d \in \{1, \dots, D\}$:

- For each topic $k \in \{1, 2, \dots\}$:
  - ○ Sample topic activation $\bar{\theta}_{kd} \sim$ dd-IBP(η);
  - ○ Sample topic distribution:
    $\Theta \sim$ dd-ICDP(α, η), $\Theta | \bar{\Theta} \sim$ DP($\alpha \bar{\Theta}$);
- For each word $i \in \{1, 2, \dots, N\}$:
  - ○ Sample topic $z_{di} \sim$ Multinomial ($\vec{\theta}_d$);
  - ○ Sample word $w_{di} \sim$ Multinomial ($\Phi_{z_{di}}$);

We define $\bar{\Theta}$ as the binary selection mask for $\Theta \in \mathbb{R}^{K \times D}$ where the dd-ICDP of topic distributions is obtained by integrating out the latent binary mask $\bar{\Theta}$:

$$\Theta \sim \text{dd-ICDP}(\alpha, \eta) = \sum_{\bar{\Theta}} p(\Theta | \bar{\Theta}) p(\bar{\Theta}), \tag{214}$$

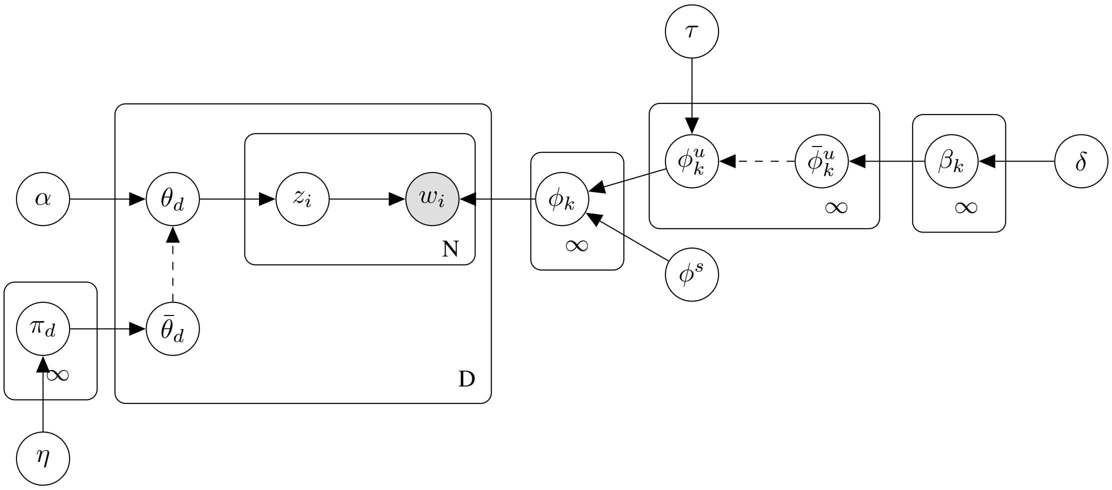


Figure 8.3: A graphical model representation for interactive Distance Dependent IBP compound Dirichlet Process (idd-ICDP)

In a similar way, we present the binary mask $\bar{\Phi}^u$ for word distributions with size $K \times V$ defined by:

$$\Phi^u \sim \text{dd-ICDP}(\delta, \tau) = \sum_{\bar{\Phi}^u} p(\Phi^u|\bar{\Phi}^u)p(\bar{\Phi}^u), \tag{215}$$

8.6 Inference

In this section, we explore the inference of each parameter that defines the proposed idd-ICDP approach. We mention that the proposed approach is the application of distance dependant Indian Buffet Process (dd-IBP) compound Dirichlet process to dual-sparse topic model in an interactive framework. For that, we determine first the set of focused topics and focused terms. Next, we update the parameters using Markov Chain Monte Carlo sampling approach.

We begin by giving the marginal likelihood associated with topics and words as follows:

$$\mathbf{Z}|\bar{\Theta} \sim \prod_d \frac{\Gamma(\alpha\bar{\theta}_{.d})}{\Gamma(\alpha\bar{\theta}_{.d}+n_{..d})} \prod_{k:\bar{\theta}_{kd}\neq 0} \frac{\Gamma(\alpha\bar{\theta}_{kd}+n_{.kd})}{\Gamma(\alpha\bar{\theta}_{kd})}, \tag{216}$$

$$\mathbf{W}|\mathbf{Z}, \bar{\Phi}^u \sim \prod_k \frac{\Gamma(\tau\bar{\phi}^u_{.k})}{\Gamma(\tau\bar{\phi}^u_{.k}+n_{.k})} \prod_{v:\bar{\phi}^u_{kv}\neq 0} \frac{\Gamma(\tau\bar{\phi}^u_{kv}+n_{vk.})}{\Gamma(\tau\bar{\phi}^u_{kv})}, \tag{217}$$

where n_{vkd} denotes the number of times word v was assigned or inherited to topic k in document d.

The notation . means the summation over the corresponding index. We mention that if $\bar{\theta}_{kd} = 0$ and $\bar{\phi}_{vk} = 0$ we will have $n_{.kd} = 0$ and $n_{vk.} = 0$, respectively.

To sample the *connectivity matrix*, the owned words b^*_v, and the owned topics c^*_k, we use the Metropolis algorithm based on the likelihood ratio. For that, we can write the joint posterior over topic assignment matrix B and document assignment matrix C:

$$\begin{aligned} p(B, b^*, \tau, \delta|\Phi^u, M, f) &\propto p(\Phi^u|B, b^*, \tau\bar{\Phi}^u)p(B|M, f) \\ &\quad p(b^*|\delta)p(\delta), \quad (218) \\ p(C, c^*, \alpha, \eta|\Theta, Q, f) &\propto p(\Theta|C, c^*, \alpha\bar{\Theta})p(C|Q, f) \\ &\quad p(c^*|\eta)p(\eta), \quad (219) \end{aligned}$$

where $p(\Phi^u|B, b^*, \tau\bar{\Phi}^u)$ and $p(\Theta|C, c^*, \alpha\bar{\Theta})$ are the likelihoods, the second terms in both equations are the priors over parameters, $p(B|M, f)$ and $p(C|Q, f)$ are the dd-ICDP priors over the connectivity matrix B and C respectively, $p(b^*|\delta)$ and $p(c^*|\eta)$ are the priors over the ownership vectors and the last terms are the priors over δ and η. We recall that $\Theta \in \mathbb{R}^{K\times D}$ represents topic proportions and $\Phi^u \in \mathbb{R}^{K\times V}$ represents the words proportions matrices. In dd-ICDP, the topic and word activation does not operate with random binary matrix yet with deterministic (many-to-one) functions on the random variables B, b^*, and C, c^* for word activation $\bar{\Phi}^u$ and topic activation $\bar{\Theta}$, respectively. We denote the deterministic function by γ and we compute the Gibbs updates as follows:

$$\begin{aligned} p(\bar{\Phi}^u|\mathbf{W}, \mathbf{Z}, M, \delta, f) &\propto p(\mathbf{W}|\mathbf{Z}, \bar{\Phi}^u)p(\bar{\Phi}^u|M, \delta, f), \quad (220) \\ p(\bar{\Theta}|\mathbf{Z}, Q, \eta, f) &\propto p(\mathbf{Z}|\bar{\Theta})p(\bar{\Theta}|Q, \eta, f), \quad (221) \\ p(z_i = k|\mathbf{W}, \mathbf{Z}^{\backslash i}, \bar{\Theta}, \Phi) &\propto p(\mathbf{W}|\mathbf{Z}, \bar{\Phi}^u, \Phi^s)p(\mathbf{Z}|\bar{\Theta}). \quad (222) \end{aligned}$$

where $\Phi = \lambda_1\Phi^u + \lambda_2\Phi^s$ represents the interactive dd-ICDP model with introducing a new variable Φ^s to denote the subjective topic-word distribution generated by human experts, λ_1, λ_2 are the weights of Φ^u and Φ^s, respectively, and

$$\begin{aligned} p(\bar{\Phi}^u|M,\delta,f) &= \sum_{(B,b^*):\gamma(B,b^*)=\bar{\Phi}} p(B,b^*|M,\delta,f) \\ &= \sum_{(B,b^*):\gamma(B,b^*)=\bar{\Phi}} p(b^*|\delta)p(B|b^*,M,f) \\ &= \sum_{(B,b^*):\gamma(B,b^*)=\bar{\Phi}} \prod_{k=1}^{K} p(\sigma_k|\delta) \prod_{k=1}^{K}\prod_{v=1}^{V} a_{kb_{vk}}, \end{aligned} \tag{223}$$

and

$$\begin{aligned} p(\bar{\Theta}|Q,\eta,f) &= \sum_{(C,c^*):\gamma(C,c^*)=\bar{\Theta}} p(C,c^*|Q,\eta,f) \\ &= \sum_{(C,c^*):\gamma(C,c^*)=\bar{\Theta}} p(c^*|\eta)p(C|c^*,Q,f) \\ &= \sum_{(C,c^*):\gamma(C,c^*)=\bar{\Theta}} \prod_{d=1}^{D} p(\mu_d|\eta) \prod_{d=1}^{D}\prod_{k=1}^{K} g_{dc_{kd}}, \end{aligned} \tag{224}$$

Various decay functions f could be defined in this approach such as:

- Constant function: $f(x) = 1$ if $d < \infty$ and $f(\infty) = 0$.
- Exponential function: $f(x) = \exp(-\beta x)$.
- Logistic function: $f(x) = 1/(1 + \exp(\beta x - \epsilon)$.

We mention that when using a constant decay function and a sequential distance matrix ($m_{kk'} = \infty$ for $k' > k$) the dd-IBP reduces to the standard IBP.

8.6.1 Sampling word activations

We update the word activations per topic using Gibbs sampling according to equation 220. Under dd-IBP approach, two topics ($k, k' \in \{1, \ldots, \kappa\}$) share a word if that word is activated for both ($\bar{\phi}^u_{vk} = \bar{\phi}^u_{vk'} = 1$) for $k \neq k'$ and a given word v. We define $\varphi_k = \sum_{v=1}^{\infty} \bar{\phi}^u_{vk}$ the number of words held by topic k and $\varphi_{kk'} = \sum_{v=1}^{\infty} \bar{\phi}^u_{vk}\bar{\phi}^u_{vk'}$ the number of words shared by topics k and k', where $k \neq k'$. The probability that $\bar{\phi}_{vk}$ is activated is given by:

$$\bar{\phi}_{k,\varkappa(v)}|\mathbb{K} \sim \text{Bernoulli}(\beta_k), \tag{225}$$

where

$$\begin{aligned} \beta_k &= \sum_{\kappa=1}^{K} p(b^*_{\varkappa(v)} = \kappa|\mathbb{K})p(\mathcal{R}_{k,\kappa,\varkappa(v)} = 1|\mathbb{K}) \\ &= y_k^{-1}/\sum_{\kappa=1}^{K} y_\kappa^{-1}p(\mathcal{R}_{k\kappa} = 1), \end{aligned} \tag{226}$$

where $\varkappa(v)$ is a uniform random permutation of $\{1, \ldots, V\}$ and $p(\mathcal{R}_{k,\kappa,\varkappa(v)} = 1|\mathbb{K}) = 1$ does not depend on v, for that we have dropped $\varkappa(v)$ in calculating the probability of the reachability. We define also the probability of activating words-sharing under the same topic as follows:

$$\bar{\phi}_{k,\varkappa(v)}, \bar{\phi}_{k',\varkappa(v)}|\mathbb{K} \sim \text{Bernoulli}(\beta_{kk'}), \tag{227}$$

where

$$\begin{aligned} \beta_{kk'} &= \sum_{\kappa=1}^{K} p(b^*_{\varkappa(v)} = \kappa|\mathbb{K})p(\mathcal{R}_{k,\kappa,\varkappa(v)} = 1, \mathcal{R}_{k',\kappa,\varkappa(v)}|\mathbb{K}) \\ &= y_k^{-1}/\sum_{\kappa=1}^{K} y_\kappa^{-1}p(\mathcal{R}_{k\kappa} = 1, \mathcal{R}_{k'\kappa} = 1), \end{aligned} \tag{228}$$

8.6.2 Sampling assignments for owned words

We update topic assignments for owned words corresponding to active features (words) using Gibbs update of the connectivity matrix. The elements of **B** are sampled according to the following probabilities:

$$\begin{aligned} p(b_{vk}|B_{-k}, \Phi^u_k, b^*, M, \tau, f) &= p(\Phi_k|B, b^*, \tau\bar{\Phi}^u_k) \\ &\quad p(b_{vk}|M, f), \end{aligned} \tag{229}$$

where B_{-k} is the connectivity matrix B excluding row k, $p(\Phi_k|B, b^*, \tau\bar{\Phi}_k^u)$ is the likelihood including only the active columns of $\bar{\Phi}$ (i.e., $\varphi_k > 0$), and $p(b_{vk}|M, f)$ is the prior given by $p(b_{vk} = k'|M, f) = a_{kk'}$. To assign b_{vk}, two possible scenarios could occur: a topic k reaches the owner of v for which the word v becomes active for topic k and for all the topics that reach k, or it does not (word v becomes inactive).

8.6.3 Sampling word ownership

In this subsection, we represent how we update the word ownership and topic assignment for newly owned words using Metropolis-Hastings algorithm. We define a new ownership vector $b_v^{*'}$ and a new connectivity matrix B' corresponding to the new allocated words going from inactive to active feature in the sampling step. We update samples from the Metropolis ratio defined by the likelihood of the new defined ownership and connectivity matrix and the likelihood given by equation 229. The sampler proceeds as follows:

(1) Sample $\sigma_k' \sim \text{Poisson}(\delta/y_k)$ for $k = 1, \ldots, K$, let $\mathbb{V}_k' = (\sum_{j<k} \sigma_j', \sum_{j\leq k} \sigma_j']$, and set the new ownership $b_v^{*'} = k$, for all $v \in \mathbb{V}_k'$.

(2) Assign $B' \leftarrow B$, for each $k = 1, \ldots, K$:

   a. If $\sigma_k' > \sigma_k$, allocate $\sigma_k' - \sigma_k$ new words to topic k.

   To insert this new words in the new connectivity matrix B', we relabel words owned by later topics via moving each column $v > \sum_{j<k} \sigma_j' + \sigma_k$ to column $v + \sigma_k' - \sigma_k$ in B'. Then, for each new word $v \in (\sum_{j<k} \sigma_j' + \sigma_k, \sum_{j\leq k} \sigma_j']$ fill in the corresponding column of B' by sampling b_{mv}' corresponding to $p(b_{vm}' = j) = a_{mj}$.

   b. If $\sigma_k' < \sigma_k$, remove $\sigma_k - \sigma_k'$ randomly selected words from topic k.

   First, we choose $\sigma_k - \sigma_k'$ words uniformly at random from $(\sum_{j<k} \sigma_j', \sum_{j\leq k} \sigma_j' + \sigma_k]$. Then, remove these columns from B' and relabel all words after the first removed word by moving the corresponding columns of B'.

(3) Compute the acceptance ratio as follows:

$$\zeta_k = \min\left[1, \frac{p(\Phi_k|B', b'^*, \tau\bar{\Phi}_k^u)}{p(\Phi_k|B, b^*, \tau\bar{\Phi}_k^u)}\right], \tag{230}$$

(4) Draw ϖ_k ~Bernoulli(ζ_k).

If $\varpi_k = 1$, set $B \leftarrow B'$ and $b^* \leftarrow b^{'*}$, otherwise keep B and b^* unchanged.

8.6.4 Sampling topic activations

Topic activations per document are sampled based on the Gibbs update given in equation 221. If two documents d, d' share the same topic k, we state that topic k is active for both ($\bar{\theta}_{kd} = \bar{\theta}_{kd'} = 1$), $d \neq d'$. The number of topics held by document d: $\vartheta_d = \sum_{k=1}^{\infty} \bar{\theta}_{kd}$ and the number of topics shared by document d and d' is $\vartheta_{dd'} = \sum_{k=1}^{\infty} \bar{\theta}_{kd}\bar{\theta}_{kd'}$ for $d \neq d'$. The probability that $\bar{\theta}_{kd}$ is activated is given by:

$$\bar{\theta}_{d,\varkappa(k)}|\mathbb{D} \sim \text{Bernoulli}(\pi_d), \tag{231}$$

where

$$\begin{aligned} \pi_d &= \sum_{\varrho=1}^{D} p(c^*_{\varkappa(k)} = \varrho|\mathbb{D}) p(\mathcal{L}_{d,\varrho,\varkappa(k)} = 1|\mathbb{D}) \\ &= h_d^{-1} / \sum_{\varrho=1}^{D} h_\varrho^{-1} p(\mathcal{L}_{d\varrho} = 1), \end{aligned} \tag{232}$$

where $\varkappa(k)$ is a uniform random permutation of $\{1, \ldots, K\}$ and $p(\mathcal{L}_{d,\varrho,\varkappa(k)} = 1|\mathbb{D}) = 1$ does not depend on k, for that we have dropped $\varkappa(k)$ in the last line. We define also the probability of activating topics-sharing under the same document as follows:

$$\bar{\theta}_{d,\varkappa(k)}, \bar{\theta}_{d',\varkappa(k)}|\mathbb{D} \sim \text{Bernoulli}(\pi_{dd'}), \tag{233}$$

where

$$\begin{aligned} \pi_{dd'} &= \sum_{\varrho=1}^{D} p(c^*_{\varkappa(k)} = \varrho|\mathbb{D}) p(\mathcal{L}_{d,\varrho,\varkappa(k)} = 1, \mathcal{L}_{d',\varrho,\varkappa(k)}|\mathbb{D}) \\ &= h_d^{-1} / \sum_{\varrho=1}^{D} h_\varrho^{-1} p(\mathcal{L}_{d\varrho} = 1, \mathcal{L}_{d'\varrho} = 1), \end{aligned} \tag{234}$$

8.6.5 Sampling assignments for owned topics

The document assignments for owned topics are updated corresponding to active topics according to the sampling of the connectivity matrix:

$$\begin{aligned} p(c_{kd}|C_{-d}, \Theta_d, c^*, Q, \alpha, f) &= p(\Theta_d|C, c^*, \alpha\bar{\Theta}_d) \\ & \quad p(c_{kd}|Q, f), \end{aligned} \tag{235}$$

where C_{-d} is the connectivity matrix C excluding row d, $p(\Theta_d|C, c^*, \alpha\bar{\Theta}_d)$ is the likelihood including only the active columns of $\bar{\Theta}$ where $\vartheta_d > 0$, and $p(c_{kd}|Q, f)$ is the prior given by $p(c_{kd} = d'|Q, f) = g_{dd'}$. As we consider the assignment of c_{kd}, one of two scenarios could occur: a document d reaches the owner of k for which the topic k becomes active for document d and for all the documents that reach d, or it does not reach the owner and the topic k becomes inactive.

8.6.6 Sampling topic ownership

We update topic ownership and document assignment for newly owned topics using the same methodology of sampling word ownership where we summarized as following:

(1) Sample $\mu'_d \sim$ Poisson(η/h_d) for $d = 1, \ldots, D$, let $\mathbb{K}'_d = (\sum_{c<d} \mu'_c, \sum_{c \leq d} \mu'_c]$, and set the new ownership $c_k^{*'} = d$, for all $k \in \mathbb{K}'_d$.

(2) Assign $C' \leftarrow C$, for each $d = 1, \ldots, D$:

a. If $\mu'_d > \mu_d$, allocate $\mu'_d - \mu_d$ new topics to document d.

b. If $\mu'_d < \mu_d$, remove $\mu_d - \mu'_d$ randomly selected topics from document d.

(3) Compute the acceptance ratio as follows:

$$\zeta_d \quad = \quad \min\left[1, \frac{p(\Theta_d|C', c'^*, \alpha\bar{\Theta}_d)}{p(\Theta_d|C, c^*, \alpha\bar{\Theta}_d)}\right], \tag{236}$$

(4) Draw ϖ_d ~Bernoulli(ζ_d).

If $\varpi_d = 1$, set $C \leftarrow C'$ and $c^* \leftarrow c^{'*}$, otherwise keep C and c^* unchanged.

8.6.7 Sampling hyperparameters

To sample the hyperparameters α and τ, we obtain the following Gibbs updates:

$$p(\alpha|\mathbf{Z}, \bar{\Theta}) \quad \propto \quad p(\mathbf{Z}|\bar{\Theta}, \alpha)p(\alpha), \tag{237}$$

$$p(\tau|\mathbf{w}, \mathbf{Z}, \bar{\Phi}^u) \quad \propto \quad p(\mathbf{w}|\mathbf{Z}, \bar{\Phi}^u, \tau)p(\tau), \tag{238}$$

where $p(\mathbf{Z}|\bar{\Theta}, \alpha)$ is given in equation 216, $p(\alpha) \propto \frac{1}{\alpha}$, $p(\mathbf{w}|\mathbf{Z}, \bar{\Phi}^u, \tau)$ is given by 217, and $p(\tau) \propto \frac{1}{\tau}$. In order to sample the δ and η hyperparameters, we draw from the following conditional distributions:

$$p(\delta|b^*, M, f) \quad \propto \quad p(\delta)\prod_{k=1}^{K}\text{Poisson}(\sigma_k; \delta/y_k), \tag{239}$$

$$p(\eta|c^*, Q, f) \quad \propto \quad p(\eta)\prod_{d=1}^{D}\text{Poisson}(\mu_d; \eta/h_d), \tag{240}$$

where σ_k and μ_d are determined by b^* and c^*, respectively. The priors on δ and η are chosen as Gamma distributions with shape parameters a_δ, a_η and scale parameters b_δ, b_η. Using the conjugate properties of Gamma and Poisson distributions, we give the conditional probability over δ and η as follows:

$$\delta|b^*, M, f \sim \text{Gamma}(a_\delta + \sum_{k=1}^{K} \sigma_k, b_\delta + \sum_{k=1}^{K} y_k^{-1}), \tag{241}$$

$$\eta|c^*, Q, f \sim \text{Gamma}(a_\eta + \sum_{d=1}^{D} \mu_d, b_\eta + \sum_{d=1}^{D} h_d^{-1}), \tag{242}$$

8.6.8 Sampling interactive topic assignments

In the estimation step of the interactive framework of the model, we compute linear combination of subjective distribution and objective distribution before sampling topic index of each word in documents. The posterior probability introduced in equation 222 is given by:

$$p(z_i = k|\mathbf{w}, \mathbf{Z}^{\setminus i}, \bar{\Theta}, \Phi) \propto \frac{\alpha + n_{.kd}^{\setminus i}}{\bar{\phi}_{.k}\beta + n_{.k}^{\setminus i}} \bar{\theta}_{kd} \left[\lambda_1(\beta + n_{vk.}^{\setminus i})\bar{\phi}_{vk}^u + \lambda_2 \phi_{vk}^s \right], \tag{243}$$

To calculate the subjective topic-word distribution, Liu et al. [280] proposed two strategy for iLDA: deterministic strategy that adjusts the objective topic-word distribution according to the degenerative probabilities completely and the stochastic strategy which takes expert confidence into account to adjust the topic-word distribution. In our model, we consider the stochastic strategy that reflect the belief of human experts to adjust the objective distribution according to their knowledge. From T most probable words in topic k that correspond to topic-word probabilities $\bar{\phi}_{kt}^u = \{p_{kt}^u, t = 1, \ldots, T\}$, human experts select T' words with degenerative probabilities $\{p_{kt'}^s \in [0, 1), t' = 1, \ldots, T'\}$. We update the subjective topic-word probabilities for the words adjusted bu human experts as follows:

$$\phi_{kT'}^s = \{p_{kt}^u \times [p_{kt'}^s]^{I(u > p_{kt'}^s)} | t \in T, t' \in T'\}, \tag{244}$$

where $I(u > p_{kt'}^s)$ is a signal function that is equal to 1 if $u > p_{kt'}^s$ and 0 otherwise, u is a stochastic variable defined to determine whether the probability of a word should be adjusted. To normalize the adjusted words and to make the sum of the probabilities of all words in a topic equals 1, we

Table 8.2: Statistics of the datasets

Dataset	Number of documents	Vocabulary size
NIPS	1,740	12,419
20 Newsgroup	18,774	60,698
Twitter	40,000	32,641
DBLP	40,190	9,393

define weighted degenerative probabilities of the remaining words in $T - T'$ as follows:

$$\begin{aligned}\phi^s_{k,-T'} &= \Bigg\{p_{-t'} \times \Big(1 + p_r \frac{p_{-t'}}{\sum_{-T'} p_{-t'}}\Big) | -t' \in T \\ &\qquad \text{and} -t' \neq T'\Bigg\}, \end{aligned} \tag{245}$$

where $p_r = \sum_{t'=1}^{T'} p_{t'} \times (1 - p^s_{kt'})$ and$\{p^s_{kt'}t' = 1, \ldots, T', T' \leq T\}$ are the probabilities for words in T'.

Then, we obtain the subjective topic-word distribution:

$$\Phi^s_k = \{\phi^s_{kT'}, \phi^s_{k,-T'}, \phi^s_{k,V-T}\}, \tag{246}$$

where $\phi_{k,V-T'}$ is the topic-word probabilities for words in $V - T'$.

8.7 Experiments

In this section, we evaluate the performance of the proposed iddICDP and ddICDP in several benchmark datasets by comparing with other related sparse topic models. Our experiments include qualitative and quantitative analysis.

8.7.1 Datasets

We consider four benchmark corpora: NIPS, 20 newsgroup, Twitter, and DBLP. Data characteristics are reported in Table 8.2. We adopt for the mentioned datasets the conventional pre-processing steps which includes tokenization, removing stop words, and removing infrequent words.

(1) **NIPS dataset**[1]: contains $1,740$ research papers collected from proceeding of Neural Information Processing Systems (NIPS) conferences (1987-1999).

(2) **20 Newsgroups**[2]: consists of $18,774$ documents categorized in 20 different newsgroups with a vocabulary of $60,698$ unique words.

(3) **Twitter**[3]: contains a collection of $40,000$ tweets from the Twitter dataset of the Democratic presidential candidates, starting in January 2019.

(4) **DBLP**[4]: contains $40,190$ documents from conference papers from three research areas: data mining/information retrieval, theoretical computer science, and computer network systems.

8.7.2 Methods for comparison

We compare iddICDP with the following state-of-the-art sparse topic models and the baseline methods:

(1) **LDA** [181]: is the classical topic model used to analyze topics from a collection of documents.

(2) **FTM** [267]: is a sparse topic model associating an IBP prior for the document-topic distribution.

(3) **DsparseTM** [268]: is a dual sparse topic model considering bounded focused topics (topic selector) and finite number of focused terms (term selector).

(4) **STC** [286]: is a sparse topical coding which introduces also the "Spike and Slab" prior by introducing Laplacian to control the sparsity of inferred representations.

(5) **LIDA** [2]: is a dual sparse topic model imposing a three-parameter IBP compound Dirichlet process on the topic and the word proportions.

[1] https://cs.nyu.edu/ roweis/data.html

[2] http://qwone.com/ jason/20Newsgroups/

[3] https://nyc3.digitaloceanspaces.com/ml-files-distro/v1/bloomberg-tweet-topics/data/tweets.csv

[4] http://www.informatik.uni-trier.de/db

8.7.3 Evaluation metrics

In the experiments, we adopt widely-used metrics to evaluate the performance of the proposed model and to compare the quality of the topics with the related models.

Topic coherence

The first evaluation metric used in this work is the point-wise mutual information (PMI) employed to measure the semantic coherence of the learned topics and known as the coherence score. For a given topic T, we consider the top-N most probable words $w_1, \ldots, w_N$, and the PMI score is defined as:

$$\text{PMI-score} = \frac{2}{N(N-1)} \sum_{1 \neq i < j \leq N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{247}$$

where $p(w_i, w_j)$ is the joint probability that both words w_i and w_j occur in a same document, $p(w_i)$ and $p(w_j)$ are the marginal probabilities of words w_i and w_j respectively.

Sparsity

Sparsity metric refers to the sparsity of topics outlined early in [266] (equation 207). The dual-sparsity defined in [268] presents a quantitative method to measure the sparsity of the topic representation of documents and the word representation of topics. We consider for our experiments the expectation of the sparsity which is conditioned on the Bernoulli parameters for document-topic and topic-term distributions:

$$\text{Sparsity-ratio}(d) = \mathbb{E}[\text{sparsity}(d)] = 1 - \pi_d \tag{248}$$

$$\text{Sparsity-ratio}(k) = \mathbb{E}[\text{sparsity}(k)] = 1 - \beta_k \tag{249}$$

Classification metrics

One of the most effective applications of topic modeling is documents classification where each document is represented with its topic distribution. Supervised topics models [287] embed topic-document distribution as the features input for SVM classifier to predict document classes. We

Table 8.3: Topic coherence (PMI) performance of all the models on four datasets with different numbers of topics and the number of topic words is equal to 15.

	20 Newsgroups	DBLP	Twitter	NIPS
Number of topics	120	15	200	30
LDA	1.336	0.622	0.562	0.623
SATM [288]	-	-	-	1.05
STC [286]	1.51	0.08	0.37	-
NSTM [289]	0.23	-	-	-
DsparseTM [268]	1.621	0.871	1.051	-
ddICDP	1.36	2.71	2.55	1.58
iddICDP	**2.28**	**2.77**	**2.68**	**1.86**

evaluate the effectiveness of the proposed model in text classification through the accuracy, recall, precision, and F1-measure for a collection of D documents:

$$\text{Accuracy} = \frac{1}{|D|} \sum_{d \in D} I(\text{Label}_d = \text{Prediction}_d), \tag{250}$$

where $I(.)$ is an indicator function, Label_d, and Prediction_d are the true label and the predicted label of document d,

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{251}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{252}$$

$$\text{F1-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{253}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

8.7.4 Quantitative analysis

We study the influence of the initial number of topics found by ddICDP on the topic coherence for 20 NG, DBLP, Twitter, and NIPS datasets in Figure 8.5. We notice that the performance of the ddICDP model are not easily affected by the initial number of topics as our proposed approach is a non-parametric Bayesian model and the number of topics is inferred through sampling the topic

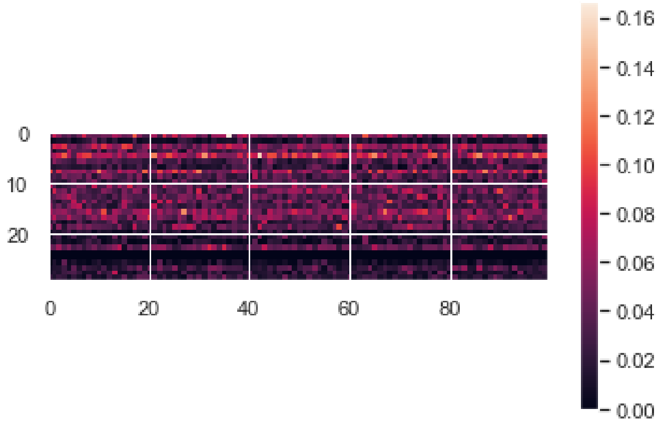


Figure 8.4: Inferred latent topics for DBLP dataset. Rows correspond to features dimensions and columns correspond to topics.

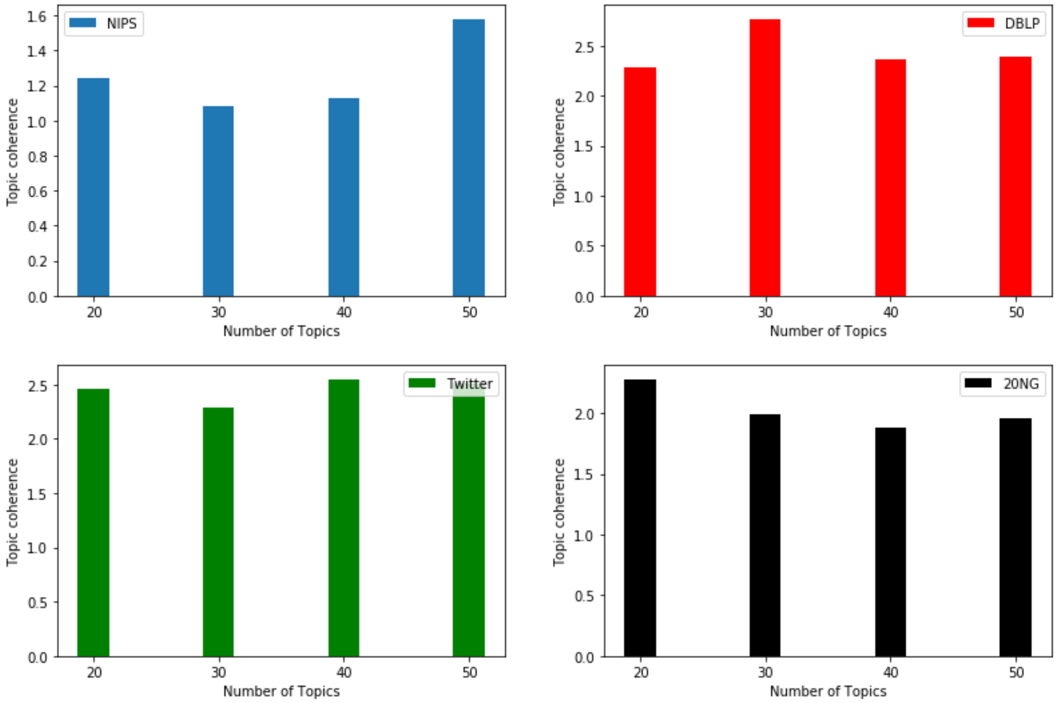


Figure 8.5: Influence of initial number of topics founded by ddICDP for different datasets

activation.

The optimal numbers of topics extracted from the data are determined with the proposed model as shown in Figure 8.4 while for the other related models the number of topics is chosen for each dataset as mentioned in Table 8.3. We compare our obtained results for the four datasets namely: 20 NG, DBLP, Twitter, and NIPS with the related models (LDA, SATM, STC, NSTM, DsparseTM). The SATM [288] is a self-aggregation based topic model proposed for short and sparse texts by aggregating short texts into long-documents while the NSTM [289] is a neural variational Sparse topic model that represents the generative process of texts with probabilistic-based approaches combined with Bidirectional LSTM for the objective of embedding contextual information over documents. We show in Table 8.3 the evaluation of the topic models for topic semantic coherence. We notice that models with sparse enhancement over topics and terms perform better than other models such as the DsparseTM, STC, SATM, and our proposed models. Our proposed methods ddICDP and iddICDP yield competitive results as compared with the other sparse topics models. We evaluate also the sparsity score corresponding to topic-word distribution as well as the average sparsity ratio of topic representation for documents for DsparseTM and the proposed models on DBLP data in Table 8.5, from which we can see that the sparsity using the iddICDP is lower than the DsparseTM and the ddICDP where the top words representative of each topics are more coherent and focused on a specified topic. We can see here the role of the interactive framework in improving the quality of words and having more representative words for each topic discovered.

8.7.5 Qualitative analysis

We illustrate in Table 8.4 an example of discovered 8 topics with top-10 representative words on NIPS dataset to capture the semantic coherence using FTM, LIDAR, and the proposed models (ddICDP, iddICDP). We compare in Table 8.5 topics inferred by DsparseTM, ddICDP, and iddICDP on DBLP dataset with their respective sparsity ratio of topic mixtures topic-document representation. From Table 8.4, we notice that all the sparse models return clean topics extracted from NIPS dataset. We can see that the proposed iddICDP can perfectly represent these topics where the irrelevant terms found in the ddICDP are updated through the interactive framework based on the subjective topic-word probabilities adjusted by human experts.

Table 8.4: Examples of discovered topics with top-10 representative words by FTM, LIDA [2], ddICDP, and iddICDP on NIPS

FTM

model	control	classifier	chip	gradient	network	learning	optimization
bayesian	controller	classification	neural	learning	system	robot	constraint
data	model	training	analog	descent	point	field	problem
parameter	learning	pattern	weight	rate	dynamic	arm	annealing
estimator	system	error	network	stochastic	attractor	model	method
variables	task	set	neuron	momentum	delay	control	objective
method	critic	class	implementation	convergence	neural	dynamic	solution
variance	forward	data	circuit	error	fixed	motor	energy
criterion	actor	mlp	digital	adaptive	stability	task	neural
selection	architecture	decision	vlsi	parameter	connection	space	point

LIDA

model	control	classifier	chip	algorithm	network	model	point
data	model	classification	neural	gradient	system	movement	problem
estimation	controller	training	weight	error	dynamic	field	function
parameter	robot	problem	bit	function	point	arm	optimization
cross	learning	class	digital	descent	attractor	trajectory	objective
bayesian	task	decision	implementation	problem	neural	control	algorithm
posterior	system	set	analog	method	equation	dynamic	method
prediction	forward	performance	hardware	convergence	dynamical	motor	annealing
validation	action	error	synapse	learning	delay	point	constraint
estimate	space	data	vlsi	local	fixed	hand	neural

ddICDP

model	figure	connector	train	algorithm	network	approach	point
train	model	data	function	data	state	control	vector
data	learn	model	model	problem	input	approximation	value
parameter	input	time	weight	example	neural	vector	predict
estimate	function	value	learn	train	parameter	distribution	weight
system	error	result	data	perform	method	target	dynamic
neural	data	system	output	test	output	function	distribution
algorithm	network	learn	neural	method	there	result	output
learn	algorithm	figure	perform	control	distribute	show	on
function	neural	input	set	such	be	case	pattern

iddICDP

model	figure	connector	train	algorithm	network	approach	predict
train	model	data	function	data	distribution	control	distribution
data	learn	model	model	problem	neural	approximation	vector
parameter	input	time	weight	example	parameter	vector	value
estimate	function	value	learn	train	output	distribution	point
system	error	result	data	perform	method	target	weight
neural	data	system	output	test	vector	function	dynamic
algorithm	network	learn	neural	method	online	result	output
learn	algorithm	figure	perform	control	system	show	model
function	neural	linear	set	target	time	case	pattern

Table 8.5: Focused topics and average sparsity ratio on DBLP data by DsparseTM, ddICDP, and iddICDP

DsparseTM		
web	time	networks
search	polynomial	control
information	algorithm	performance
semantic	linear	analysis
content	computing	traffic
user	algorithms	atm
mining	systems	packet
knowledge	computation	network
(k) 0.9927	(k) 0.9963	(k) 0.9966
(d) 0.98873	(d) 0.9010	(d) 0.9210
ddICDP		
web	space	network
framework	search	cluster
field	algorithm	classifier
global	shape	feature
function	problem	coordinate
form	system	base
filter	structure	real
feature	inform	multi
(k) 0.9666	(k) 0.9333	(k) 0.9333
(d) 0.992	(d) 0.996	(d) 0.996
iddICDP		
web	scale	network
edge	structure	cluster
network	model	transform
depth	base	wavelet
detect	database	rule
outlier	bayesian	random
multimedia	statistic	sequence
reason	sequence	dynamic
(k) 0.900	(k) 0.900	(k) 0.900
(d) 0.99	(d) 0.99	(d) 0.99

8.7.6 Sentiment Analysis

Sentiment analysis is the process of mining information on social media to identify data into positive, negative, or neutral polarities. Such polarities concern opinions, reviews, attitudes, recommendations, and feelings expressed in text. This process has tremendously dispersed to include opinion mining, client's reviews, and more which becomes of great interest for several applications such as marketing, business analysis, and presidential election. Hence, much work has focused on extracting information for sentiment analysis [290, 291]. Traditional machine learning algorithms use generally the bag-of-words as the textual representation. However, the bag-of-words structure ignore the semantic information of textual data. In addition, detecting the sentiment polarities depend on topics which makes sentiment analysis an active area in Natural Language Processing (NLP). In this regard, LDA was applied to evaluate topic quality and opinions polarities on Twitter in several research works [292, 293, 294]. In this work, we consider the proposed ddICDP and iddICDP as supervised topic modeling techniques to classify documents into positive or negatives categories.

In this experiment, we use the sentiment analysis dataset: Sentiment Strength Twitter (SentiStrength) [5] introduced in [224]. The dataset contains $4,242$ tweets annotated initially into a positive strength range and negative strength. Following, Saif et al. [222] have re-annotated the labels into $1,252$ positive and $1,037$ negative tweets.

We compare the results obtained by our algorithms with SVM and supervised LDA in Table 8.6. We notice the outperformance of the proposed models comapred to SVM and sLDA when considering the accuracy, recall, precision, and F1 measure metrics. It is noteworthy to mention that the accuracy obtained by SVM gives better result than sLDA while it gives the least scores for all the other metrics. When comparing the performance of the ddICDP and the interactive one for the classification, we mention that their results are slightly different and they perform very similarly.

[5] Available at http://sentistrength.wlv.ac.uk/documentation/

Table 8.6: Classification results on SentiStrength tweets

Models	Accuracy	Recall	Precision	F1-measure
SVM	62.39	55.43	76.64	64.33
sLDA	59.78	53.42	98.77	69.34
ddICDP	65.42	58.44	81.02	71.86
iddICDP	68.47	61.01	82.56	74.10

8.8 Conclusion

This chapter introduces an interactive distance dependent IBP compound Dirichlet process (iddICDP), a nonparametric Bayesian model for topic modeling where the number of topics and terms are unbounded. The model allows to alleviate the sparsity of topic mixtures and topic-word distributions using the "Spike and Slab" process and employs a distance dependant IBP approach which shares dependency and non-exchangeable structure of text data. We generalize the dd-IBP to dd-ICDP that combines properties from the HDP and the IBP and integrates an interactive framework using human experts knowledge. We introduce a Markov Chain Monte Carlo sampler with Metropolis-Hastings algorithm for the idd-ICDP and evaluate its robustness on real-world data. Experiments on a variety of textual data demonstrates that the new type of sparse topic models we propose better fit non-exchangeable correlated sparse data in terms of point-wise mutual information and classification scores. Due to the effectiveness of our model structure, the iddICDP can be integrated with online inference procedure, particularly, for understanding online human emotions. Additionally, to address sparsity in topic models, various works have proposed stochastic variational inference on large scale data. In this regard, we anticipate that iddICDP can can be inferred also by stochastic variational Gibbs sampling algorithm.

Chapter 9

Conclusions and Future Research Directions

This thesis investigates the different challenges of count data that occur in emotion recognition and sentiment analysis. In this chapter, we summarize the main contributions of this dissertation and highlight future research directions.

9.1 Conclusions

Initially, we have proposed a reparametrization of the generalized Dirichlet multinomial (GDM) and we have computed the exact Fisher information matrix based on Beta-binomial probability density function. Validating the proposed approach on three real-world applications including three different modalities of count data: text, dialogue, and images, we have proved the benefits of taking into consideration the dependence and correlation of the feature vectors. We have demonstrated the robustness and high efficiency of the proposed model throughout all the conducted experiments as compared with existing state-of-the-art methods. Next, we have focused on multinomial-based models and a novel smoothed Dirichlet multinomial (SDM) distribution has been proposed. In this contribution, we have smoothed the count vectors to define a new distribution on a smoothed simplex and we have proved that smoothed Dirichlet is a conjugate prior to the multinomial distribution. The proposed SDM has shown the outperformance with regards to the other related multinomial-based models. The promising results on psychology analysis, pain detection, and depression on social media have proved the efficiency of the proposed algorithm in terms of accuracy scores, memory occupied, and the time of execution. Further, because of the limitations of Dirichlet distribution, we have considered the generalized Dirichlet that has more general properties and robustness in describing proportional data. We have proposed the smoothed generalized Dirichlet (SGD) distribution defined on a smoothed simplex using Jelinek-Mercer smoothing approach. In this research work, novel smoothed probabilistic approaches based on SGD have been proposed: SGD mixtures, SGD-KL, SGD-Fisher, SGD-Bhattacharyya, SGDM, and TSGDM. We have introduced a maximum likelihood approach for parameters learning and two clustering algorithms including mixture modeling and agglomerative-based geometrical information (Kulback-Leibler, Fisher information, Bhattacharyya distance). These approaches have been applied on detecting tweets emotions related to disaster and estimating pain intensity. From experimental results, we have shown the superior efficiency of the smoothed models and their robustness in modeling texts and images which have addressed the different challenges of count data. In another work, we have addressed the challenges of proportional data where we have proposed two latent topic modeling approaches based on smoothed Beta-Liouville distribution (SBL). The key idea of the first model, latent SBL, is to represent topic mixtures using SBL distributions. Next, the second model, SBL Emotion Term model, has been based on Bayesian folding-in and SBL kernels for PLSI. These two latent topic models have been able to tackle the textual data challenges through estimating new unknown fairy tales stories. From the perspective of affect recognition, SBL Emotion Term model has successfully recognized affect states from body and face video sequences within a bimodal framework. In another research work, we have addressed the sparsity problem throughout an information retrieval framework based on smoothed Scaled Dirichlet distribution. The proposed approach has been considered for the purpose of classifying sentiment tweets which has achieved promising results compared with the related-works approaches. Following, we have introduced sparse adaptive hierarchical priors for the multinomial Naive Bayes classifier. We have considered first the generalized Dirichlet as a prior which reduces the shortcomings of Bayesian approaches. The proposed model has exploited the knowledge of vocabularies structure and has been considered for predicting emotions in German and English poetry. It has achieved competitive results and has shown the effectiveness in predicting emotions from textual data. Second, the Beta-Liouville has been presented as an alternative prior for the multinomial Naive Bayes with vocabulary knowledge. Experimental results on emotion intensity analysis and hate speech detection have shown the ability to tackle the problem of sparse documents with comparison to the other related Bayes classifiers. In the last chapter, we have presented a new sparse topic model using the “Spike and slab” prior which allows to alleviate the sparsity of topics and terms. The proposed model iddICDP has been based on nonparametric Bayesian model and a distance dependent IBP approach which address the dependency and non-exchangeable structure of data. We have considered also an interactive framework for the proposed sparse topic model that has proved to be a better fit for non-exchangeable sparse model in terms of topic coherence and sentiment analysis.

9.2 Future directions

Regarding the promising results of the proposed smoothed models, we advocate seeking to analyze emotion states from other modalities such as voice signals. In addition, a potential future work can be devoted to developing a multimodal emotion recognition framework which combines all the different modalities including text, images, videos, and signal. Our work could open different scopes for exploration. Indeed, it seems worth exploring the different proposed models for other interesting applications in biological sciences, business management, and also in medicine. Moreover, the smoothed Beta-Liouville distribution and smoothed Scaled Dirichlet could be considered as conjugate priors to the multinomial distribution which could present alternative models for tackling the problems of word burstiness and sparsity challenge. Further, due to the effectiveness of our sparse topic model structure, the iddICDP can be integrated with online inference procedure, particularly, for understanding online human emotions. Additionally, to address sparsity issue, various works have proposed stochastic variational inference on large scale data. In this regard, we anticipate that the different smoothed probabilistic-based models can be inferred also by stochastic variational inference.

Appendix A

Proof of Smoothed Generalized Dirichlet equations (Chapter 4)

A.1 Proof of SGD parameters estimation

We have the log-likelihood function:

$$\begin{aligned}
\mathcal{L}(\mathcal{X}|\vec{\alpha},\vec{\beta}) &= \sum_{i=1}^{N}\sum_{d=1}^{D-1}\log\Big[\frac{(\alpha_d+\beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}}\Big] \\
&+ \log\Big[(x_d^s)^{\alpha_d-1}\big(1-\sum_{k=1}^{d}x_k^s\big)^{\gamma_d}\Big] \\
&= (\alpha_d+\beta_d)\log(\alpha_d+\beta_d)-\alpha_d\log\alpha_d \\
&- \beta_d\log\beta_d+(\alpha_d-1)\log(x_d^s) \\
&+ (\beta_d-\alpha_{d+1}-\beta_{d+1})\log\big(1-\sum_{k=1}^{d}x_k^s\big)
\end{aligned}$$

$$\frac{\partial\mathcal{L}(\mathcal{X}|\vec{\alpha},\vec{\beta})}{\partial\alpha_d} = 0$$

Deriving the log-likelihood with regards to the α_d, we obtain

$$\log(\alpha_d + \beta_d) + 1 - \log \alpha_d - 1 + \log(x_d^s) \quad = \quad 0$$

$\Leftrightarrow$

$$\log \frac{\alpha_d + \beta_d}{\alpha_d} = \log \frac{1}{x_d^s}$$

$\Leftrightarrow$

$$\alpha_d = \beta_d \, \frac{x_d^s}{1 - x_d^s}$$

Second, we derive the log-likelihood function with respect to β_d:

$$\frac{\partial \mathcal{L}(\mathcal{X}|\vec{\alpha}, \vec{\beta})}{\partial \beta_d} \quad = \quad 0$$

$\Leftrightarrow$

$$\log(\alpha_d + \beta_d) + 1 - \log \beta_d - 1 + \log \big(1 - \sum_{k=1}^{d} x_k^s\big) \quad = \quad 0$$

$\Leftrightarrow$

$$\log \frac{\alpha_d + \beta_d}{\beta_d} = \log \frac{1}{1 - \sum_{k=1}^{d} x_k^s}$$

$$\beta_d = \alpha_d \, \frac{1 - \sum_{k=1}^{d} x_k^s}{\sum_{k=1}^{d} x_k^s}$$

A.2 Proof of Geometrical information distances

A.2.1 Kulback-leibler divergence

Having the following properties for KL for two distributions that belongs to the exponential family:

$$\begin{aligned}\mathcal{K}(p(\vec{X}|\Theta), p(\vec{X}|\Theta')) &= F(\Theta) - F(\Theta') \\ &+ [G(\Theta) - G(\Theta')]^{tr} E_{\Theta}[T(X)]\end{aligned}$$

where

$$E_{\Theta}[T(X)] = -F'(\Theta)$$

$$E_{\Theta}[\log x_1^s] = \frac{-\partial F(\Theta)}{\alpha_1} = \log \alpha_1 - \log(\alpha_1 + \beta_1)$$

$$E_{\Theta}[\log x_d^s - \log(1 - \sum_{k=1}^{d-1} x_k^s)] = \frac{-\partial F(\Theta)}{\alpha_d} = \log \frac{\alpha_d}{\alpha_d + \beta_d}$$

$$E_{\Theta}[\log(1 - x_1^s)] = \frac{-\partial F(\Theta)}{\beta_1} = \log \beta_1 - \log(\alpha_1 + \beta_1)$$

$$\begin{aligned}E_{\Theta}[\log(1 - \sum_{k=1}^{d} x_k^s) - \log(1 - \sum_{k=1}^{d-1} x_k^s)] &= \frac{-\partial F(\Theta)}{\beta_d} \\ &= \log \frac{\beta_d}{\alpha_d + \beta_d}\end{aligned}$$

A.2.2 Bhattacharya

$$
\begin{aligned}
\mathcal{B} &= \frac{1}{2}F(\Theta_1) + \frac{1}{2}F(\Theta_2) - F(\frac{1}{2}\Theta_1 + \frac{1}{2}\Theta_2) \\
&= \frac{1}{2}\sum_{d=1}^{D-1}(\alpha_{d1} + \beta_{d1})\log(\alpha_{d1} + \beta_{d1}) - \alpha_{d1}\log\alpha_{d1} \\
&- \beta_{d1}\log\beta_{d1} + (\alpha_{d2} + \beta_{d2})\log(\alpha_{d2} + \beta_{d2}) - \alpha_{d2}\log\alpha_{d2} \\
&- -\beta_{d2}\log\beta_{d2} - \frac{1}{2}\sum_{d=1}^{D-1}((\alpha_{d1} + \beta_{d1}) + (\alpha_{d2} + \beta_{d2})) \\
&\quad \log(\frac{\alpha_{d1} + \beta_{d1}}{2} + \frac{\alpha_{d2} + \beta_{d2}}{2}) - (\alpha_{d1} + \alpha_{d2}) \\
&\quad \log(\frac{\alpha_{d1} + \alpha_{d2}}{2}) - (\beta_{d1} + \beta_{d2})\log(\frac{\beta_{d1} + \beta_{d2}}{2})
\end{aligned}
$$

A.3 Proof of Conjugate Prior of SGD distribution

The joint distribution of $\vec{X}$ and $\vec{P}$ ($p(\vec{X}|\vec{P})$ is a multinomial distribution and the prior $p(\vec{P}|\vec{\alpha},\vec{\beta})$ is a Smoothed Generalized Dirichlet) given by:

$$
\begin{aligned}
p(\vec{X}|\vec{\alpha},\vec{\beta}) &= p(\vec{X}|\vec{P})p(\vec{P}|\vec{\alpha},\vec{\beta}) \\
&= \frac{|x|!}{\prod_{d=1}^{D} x_d!}\prod_{d=1}^{D-1}\frac{(\alpha_d + \beta_d)^{\alpha_d+\beta_d}}{\alpha_d^{\alpha_d}\beta_d^{\beta_d}} \\
&\quad \prod_{d=1}^{D}(P_d^s)^{x_d}\prod_{d=1}^{D-1}(P_d^s)^{\alpha_d-1}\big(1 - \sum_{k=1}^{d}P_k^s\big)^{\gamma_d}
\end{aligned}
$$

It comes to calculate the following integral:

$$\int_{\vec{P^s}} \prod_{d=1}^{D} (P_d^s)^{x_d+\alpha_d-1} \big(1 - \sum_{k=1}^{d} P_k^s\big)^{\gamma_d} d\vec{P^s}$$

$$= \int_{\vec{P^s}} (P_1^s)^{\alpha_1+x_1-1}(1-(P_1^s))^{\beta_1-\alpha_2-\beta_2}\dots$$

$$(P_D^s)^{x_D}(1-P_1^s-\dots P_{D-1}^s)^{x_D} d\vec{P^s}$$

$$= \int_{\vec{P^s}} (P_1^s)^{\alpha_1+x_1-1}(P_2^s)^{\alpha_2+x_2-1}(P_{D-1}^s)^{\alpha_{D-1}+x_{D-1}-1}$$

$$(1-(P_1^s))^{(\beta_1+x_2+\dots+x_D)-(\alpha_2+x_2)-(\beta_2+x_3+\dots+x_D)}$$

$$(1-(P_1^s)-(P_2^s))^{(\beta_2+x_3+\dots+x_D)-(\alpha_3+x_3)-(\beta_3+x_4+\dots+x_D)}\dots$$

$$(1-(P_1^s)-\dots-(P_{D-1}^s))^{\beta_{D-1}+x_D-1} d\vec{P^s}$$

$$= \int_{\vec{P^s}} \prod_{d=1}^{D} (P_d^s)^{\alpha'_d-1} \big(1 - \sum_{k=1}^{d} P_k^s\big)^{\gamma'_d} d\vec{P^s}$$

where $\alpha'_d = \alpha_d + x_d$, $\gamma'_d = \beta'_d - \alpha'_{d+1} - \beta'_{d+1}$, and $\beta'_d = \beta_d + \sum_{k=d+1}^{D} x_k$.

A.4 Approximation of SGDM

For small positives values $\alpha_d << 1$; $\beta_d << 1$, $d = 1, \dots, D-1$, we have the following properties:

$$\lim_{\alpha_d \to 0} \alpha_d^{\alpha_d} = e^{\alpha_d \log \alpha_d} = 1$$

$$\lim_{\beta_d \to 0} \beta_d^{\beta_d} = e^{\beta_d \log \beta_d} = 1$$

$$\lim_{(\alpha_d, \beta_d) \to 0} (\alpha_d + \beta_d)^{\alpha_d+\beta_d} = e^{(\alpha_d+\beta_d)\log(\alpha_d+\beta_d)} = 1$$

After first setting the above approximation, we optimize the product of α'_d and β'_d in the pdf of SGDM by dint of Taylor series expansion around zero values.

We have:

$$\begin{aligned}\log(\alpha'_d) &= \log(\alpha_d + x_d) = \log(x_d(1 + \frac{\alpha_d}{x_d})) \\ &= \log x_d + \log(1 + \frac{\alpha_d}{x_d}) \\ &\underset{0}{\approx} \log x_d + [\frac{\alpha_d}{x_d} + \epsilon(\alpha_d^2)]\end{aligned}$$

which results in:

$$\begin{aligned}(\alpha_d + \beta_d) &\underset{0}{\approx} 1 + (\alpha_d + x_d)\Big(\log x_d + \frac{\alpha_d}{x_d} + \epsilon(\alpha_d^2)\Big) \\ &\underset{0}{\approx} 1 + x_d \log(x_d)\end{aligned}$$

for β'_d, we approximate using the same methodology where the 1st order Taylor expansion is as follows

$$\begin{aligned}(\beta_d + Z_{d+1})^{\beta_d + Z_{d+1}} &\underset{0}{\approx} 1 + (\beta_d + Z_{d+1})\Big(\log Z_{d+1} \\ &+ \frac{\beta_d}{Z_{d+1}} + \epsilon(\beta_d^2)\Big) \\ &\underset{0}{\approx} 1 + Z_{d+1} \log Z_{d+1}\end{aligned}$$

where $Z_{d+1} = \sum_{k=d+1}^{D} x_k$.

Now, we assume that α_d and β_d are both very close to zero values simultaneously when we are calculating the following formula:

$$\begin{aligned}(\alpha'_d + \beta'_d)^{\alpha'_d + \beta'_d} &= (\alpha_d + \beta - d + Z_d)^{\alpha_d + \beta_d + Z_d} \\ &= e^{(\alpha_d + x_d)\log(\alpha_d + \beta_d + Z_d)} \\ &\quad\; e^{(\beta_d + Z_{d+1})\log(\alpha_d + \beta_d + Z_d)} \\ &\underset{0}{\approx} \Big(1 + x_d \log(\beta_d + Z_d) + \epsilon(\alpha_d^2)\Big) \\ &\quad\; \Big(1 + Z_{d+1} \log(\alpha_d + Z_d) + \epsilon(\beta_d^2)\Big)\end{aligned}$$

where $Z_d = x_d + Z_{d+1}$

Appendix B

Proof of Smoothed Scaled Dirichlet (Chapter 6)

B.1 Calculation of first order derivatives

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \alpha_v} &= \sum_{i=1}^{N} \Big(\log \beta_v - (\log(\alpha_v) + 1) \\ &+ \log(w_{iv}^s) + \log \alpha_+ + 1 - \log(\sum_{v=1}^{D} \beta_v w_{iv}^s \Big)\end{aligned} \tag{254}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_v} = 0 \tag{255}$$

$$\iff$$

$$\sum_{i=1}^{N} \log \alpha_v = \sum_{i=1}^{N} \Big(\log \beta_v + \log(w_{iv}^s) - \log(\sum_{v=1}^{D} \beta_v w_{iv}^s \Big) \tag{256}$$

$$\hat{\alpha}_v = \sum_{i=1}^{N} \beta_v \frac{w_{iv}^s}{\sum_{v=1}^{V} \beta_v w_{iv}^s} \tag{257}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_v} \quad = \quad \sum_{i=1}^{N} \Big(\alpha_v \frac{1}{\beta_v} - \alpha_+ \frac{w_{vi}^s}{\sum_{v=1}^{D} \beta_v w_{iv}^s} \Big) \tag{258}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_v} \quad = \quad 0 \tag{259}$$

$$\Longleftrightarrow$$

$$\sum_{i=1}^{N} \frac{1}{\beta_v} \quad = \quad \sum_{i=1}^{N} \Big(\frac{\alpha_+}{\alpha_v} \frac{w_{vi}^s}{\sum_{v=1}^{D} \beta_v w_{iv}^s} \Big) \tag{260}$$

$$\hat{\beta}_v = \sum_{i=1}^{N} \frac{\alpha_v}{\alpha_+} \frac{\sum_{v=1}^{V} \beta_v w_{iv}^s}{w_{iv}^s} \tag{261}$$

Bibliography

[1] Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'20)*, Marseille, France, May 2020. European Language Resources Association (ELRA).

[2] Cedric Archambeau, Balaji Lakshminarayanan, and Guillaume Bouchard. Latent ibp compound dirichlet allocation. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):321–333, 2014.

[3] Slava M Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59, 1996.

[4] Kenneth W Church and William A Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[5] John Scott Shonkwiler and W Douglass Shaw. Hurdle count-data models in recreation demand analysis. *Journal of Agricultural and Resource Economics*, 21:210–219, 1996.

[6] David C Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547, 1994.

[7] Andreas Lindén and Samu Mäntyniemi. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7):1414–1421, 2011.

[8] Pushpa L Gupta, Ramesh C Gupta, and Ram C Tripathi. Analysis of zero-adjusted count data. *Computational Statistics & Data Analysis*, 23(2):207–218, 1996.

[9] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[10] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys*, 34(1):1–47, 2002.

[11] Fatma Najar and Nizar Bouguila. Happiness analysis with fisher information of dirichlet-multinomial mixture model. In *Canadian Conference on Artificial Intelligence*, pages 438–444. Springer, 2020.

[12] Fatma Najar and Nizar Bouguila. Exact fisher information of generalized dirichlet multinomial distribution for count data modeling. *Information Sciences*, 586:688–703, 2022.

[13] Fatma Najar and Nizar Bouguila. Jointly smoothing word embedding and text representation. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 282–289. IEEE, 2021.

[14] Fatma Najar and Nizar Bouguila. Image categorization using agglomerative clustering based smoothed dirichlet mixtures. In *International Symposium on Visual Computing*, pages 27–38. Springer, 2020.

[15] Fatma Najar and Nizar Bouguila. Emotion recognition: A smoothed dirichlet multinomial solution. *Engineering Applications of Artificial Intelligence*, 107:104542, 2022.

[16] Fatma Najar and Nizar Bouguila. Smoothed generalized dirichlet: a novel count data model for detecting emotional states. *IEEE Transactions on Artificial Intelligence*, pages 1–1, 2021.

[17] Fatma Najar and Nizar Bouguila. Sparse generalized dirichlet prior based bayesian multinomial estimation. In *International Conference on Advanced Data Mining and Applications*, pages 177–191. Springer, 2022.

[18] Fatma Najar and Nizar Bouguila. Sparse document analysis using beta-liouville naive bayes with vocabulary knowledge. In *International Conference on Document Analysis and Recognition*, pages 351–363. Springer, 2021.

[19] Joseph M Hilbe. *Modeling count data*. Springer, 2011.

[20] A Colin Cameron and Pravin K Trivedi. Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of applied econometrics*, 1(1): 29–53, 1986.

[21] Charu C Aggarwal and ChengXiang Zhai. An introduction to text mining. In *Mining text data*, pages 1–10. Springer, 2012.

[22] Rasmus E Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552. ACM, 2005.

[23] W Duncan Wadsworth, Raffaele Argiento, Michele Guindani, Jessica Galloway-Pena, Samuel A Shelburne, and Marina Vannucci. An integrative bayesian dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC bioinformatics*, 18(1):94, 2017.

[24] Jian Zhang, Zoubin Ghahramani, and Yiming Yang. A probabilistic model for online document clustering with application to novelty detection. In *Advances in neural information processing systems*, pages 1617–1624, 2005.

[25] Yanping Xiao, Chuang Lin, Yixin Jiang, Xiaowen Chu, and Xuemin Shen. Reputation-based qos provisioning in cloud computing via dirichlet multinomial model. In *2010 IEEE International Conference on Communications*, pages 1–5. IEEE, 2010.

[26] Nizar Bouguila and Djemel Ziou. Unsupervised learning of a finite discrete mixture: Applications to texture modeling and image databases summarization. *Journal of Visual Communication and Image Representation*, 18(4):295–309, 2007.

[27] Nizar Bouguila. Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474, 2008.

[28] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[29] Nuha Zamzami and Nizar Bouguila. Deriving probabilistic svm kernels from exponential family approximations to multivariate distributions for count data. In *Mixture Models and Applications*, pages 125–153. Springer, 2020.

[30] Nuha Zamzami and Nizar Bouguila. Consumption behavior prediction using hierarchical bayesian frameworks. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, pages 31–34. IEEE, 2018.

[31] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[32] Mordechai Segal and Ehud Weinstein. A new method for evaluating the log-likelihood gradient, the hessian, and the fisher information matrix for linear dynamic systems. *IEEE Transactions on Information Theory*, 35(3):682–687, 1989.

[33] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[34] André Klein, Guy Mélard, and Peter Spreij. On the resultant property of the fisher information matrix of a vector arma process. *Linear algebra and its applications*, 403:291–313, 2005.

[35] Zhengming Li, Zheng Zhang, Jie Qin, Zhao Zhang, and Ling Shao. Discriminative fisher embedding dictionary learning algorithm for object recognition. *IEEE transactions on neural networks and learning systems*, 2019.

[36] Raimund J Ober, Qiyue Zou, and Zhiping Lin. Calculation of the fisher information matrix for multidimensional data sets. *IEEE Transactions on Signal Processing*, 51(10):2679–2691, 2003.

[37] Peng Tang, Xinggang Wang, Baoguang Shi, Xiang Bai, Wenyu Liu, and Zhuowen Tu. Deep fishernet for image classification. *IEEE transactions on neural networks and learning systems*, 2018.

[38] N. Bouguila. A data-driven mixture kernel for count data classification using support vector machines. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, pages 26–31, Oct 2008.

[39] Nagaraj K Neerchal and Jorge G Morel. Large cluster results for two parametric multinomial extra variation models. *Journal of the American Statistical Association*, 93(443):1078–1087, 1998.

[40] Xuefeng Bai, Tiejun Zhang, Chuanjun Wang, Ahmed A Abd El-Latif, and Xiamu Niu. A fully automatic player detection method based on one-class svm. *IEICE Transactions on Information and Systems*, 96(2):387–391, 2013.

[41] Nuha Zamzami and Nizar Bouguila. A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognition*, 95:36 – 47, 2019. ISSN 0031-3203.

[42] Nuha Zamzami and Nizar Bouguila. Model selection and application to high-dimensional count data clustering. *Applied Intelligence*, 49(4):1467–1488, Apr 2019. ISSN 1573-7497.

[43] Sudhir R Paul, Uditha Balasooriya, and Tathagata Banerjee. Fisher information matrix of the dirichlet-multinomial distribution. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(2):230–236, 2005.

[44] Wentao Fan, Nizar Bouguila, Ji-Xiang Du, and Xin Liu. Axially symmetric data clustering through dirichlet process mixture models of watson distributions. *IEEE transactions on neural networks and learning systems*, 30(6):1683–1694, 2018.

[45] Fatma Najar, Sami Bourouis, Rula Al-Azawi, and Ali Al-Badi. *Online Recognition via a Finite Mixture of Multivariate Generalized Gaussian Distributions*, pages 81–106. Springer International Publishing, Cham, 2020.

[46] Fatma Najar, Nuha Zamzami, and Nizar Bouguila. Fake news detection using bayesian inference. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 389–394, July 2019.

[47] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[48] Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[49] Naonori Ueda and Ryohei Nakano. Deterministic annealing em algorithm. *Neural Networks*, 11(2):271 – 282, 1998.

[50] Rohan A Baxter and Jonathan J Oliver. Finding overlapping components with mml. *Statistics and Computing*, 10(1):5–16, 2000.

[51] Ahmed S Alghamdi, Kemal Polat, Abdullah Alghoson, Abdulrahman A Alshdadi, and Ahmed A Abd El-Latif. Gaussian process regression (gpr) based non-invasive continuous blood pressure prediction method from cuff oscillometric signals. *Applied Acoustics*, 164: 107256, 2020.

[52] Alexander Franz. Independence assumptions considered harmful. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL, pages 182–189, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[53] Stephen M Scariano and James M Davenport. The effects of violations of independence assumptions in the one-way anova. *The American Statistician*, 41(2):123–129, 1987.

[54] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[55] Tzu-Tsung Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97(2-3):165–181, 1998.

[56] Nizar Bouguila and Djemel Ziou. A powerful finite mixture model based on the generalized dirichlet distribution: unsupervised learning and applications. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 280–283. IEEE, 2004.

[57] Nizar Bouguila. Deriving kernels from generalized dirichlet mixture models and applications. *Information Processing & Management*, 49(1):123–137, 2013.

[58] David Lindley and Calyampudi R. Rao. Advanced statistical methods in biometric research. 1953.

[59] World Health Organization et al. Mental health action plan 2013-2020. 2013.

[60] Joseph Prusa, Taghi M Khoshgoftaar, and Amri Napolitano. Utilizing ensemble, data sampling and feature selection techniques for improving classification performance on tweet sentiment data. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 535–542. IEEE, 2015.

[61] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. In *11th International Conference on Language Resources and Evaluation, LREC-2018*, pages 1597–1601, Miyazaki, Japan, 05 2018. European Language Resources Association.

[62] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM ’13, pages 223–232, New York, NY, USA, 2013. ACM.

[63] Kenneth W Church and William A Gale. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190, 1995.

[64] Marina Meilă and David Heckerman. An experimental comparison of model-based clustering methods. *Machine learning*, 42(1-2):9–29, 2001.

[65] Alexander Franz. Independence assumptions considered harmful. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 182–189, 1997.

[66] Rainer Winkelmann. Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, 13(4):467–474, 1995.

[67] Slava M Katz. Distribution of content words and phrases in text and language modelling. *Natural language engineering*, 2(1):15–59, 1996.

[68] Dimitris Margaritis and Sebastian Thrun. A bayesian multiresolution independence test for continuous variables. *arXiv preprint arXiv:1301.2292*, 2013.

[69] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.

[70] James E Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962.

[71] Sonia Migliorati, Gianna Serafina Monti, and Andrea Ongaro. E–m algorithm: an application to a mixture model for compositional data. In *Proceedings of the 44th scientific meeting of the italian statistical society*, 2008.

[72] Robin Hankin. A generalization of the dirichlet distribution. *Journal of Statistical Software*, 33(11):1–18, 2010.

[73] Robert H Lochner. A generalized dirichlet distribution in bayesian life testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 37(1):103–113, 1975.

[74] Franck Barthe, Fabrice Gamboa, Li-Vang Lozada-Chang, and Alain Rouault. Generalized dirichlet distributions on the ball and moments. *arXiv preprint arXiv:1002.1544*, 2010.

[75] Nuha Zamzami and Nizar Bouguila. Text modeling using multinomial scaled dirichlet distributions. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 69–80. Springer, 2018.

[76] Gianna Monti, Glòria Figueras, and Vera Pawlowsky-Glahn. Notes on the scaled dirichlet distribution. *John Wiley & Sons, Chichester*, pages 128–138, 07 2011.

[77] Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

[78] BD Sivazlian. On a multivariate extension of the gamma and beta distributions. *SIAM Journal on Applied Mathematics*, 41(2):205–209, 1981.

[79] Tzu-Tsung Wong. Alternative prior assumptions for improving the performance of naïve bayesian classifiers. *Data Mining and Knowledge Discovery*, 18(2):183–213, 2009.

[80] Nizar Bouguila. Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198, 2010.

[81] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM New York, NY, USA, 2017.

[82] R Nallapati, T Minka, and S Robertson. The smoothed-dirichlet distribution: a new building block for generative models. *CIIR Technical Report*, April 2007.

[83] Bin Zhu, Xin Guo, Kenneth Barner, and Charles Boncelet. Automatic group cohesiveness detection with multi-modal features. In *2019 International Conference on Multimodal Interaction*, pages 577–581, 2019.

[84] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019.

[85] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019.

[86] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10143–10152, 2019.

[87] Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, and Fadi Dornaika. Multi-label, multi-task cnn approach for context-based emotion recognition. *Information Fusion*, 2020.

[88] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[89] Ramesh Nallapati. *The smoothed dirichlet distribution: Understanding cross-entropy ranking in information retrieval*. PhD thesis, University of Massachusetts Amherst Dept Of Computer Science, July 2006.

[90] Nizar Bouguila and Djemel Ziou. Unsupervised learning of a finite discrete mixture model based on the multinomial dirichlet distribution: Application to texture modeling. In *Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems,PRIS 2004, In conjunction with ICEIS 2004, Porto, Portugal, April 2004*, pages 118–127, 2004.

[91] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, pages 1–10, 2013.

[92] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

[93] Liu Yi Lin, Jaime E Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason B Colditz, Beth L Hoffman, Leila M Giles, and Brian A Primack. Association between social media use and depression among us young adults. *Depression and anxiety*, 33(4):323–331, 2016.

[94] Dilara Torunoğlu, Gürkan Telseren, Özgün Sağtürk, and Murat C Ganiz. Wikipedia based semantic smoothing for twitter sentiment classification. In *2013 IEEE INISTA*, pages 1–5. IEEE, 2013.

[95] Ray R Larson. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4):852–853, 2010.

[96] Naonori Ueda and Ryohei Nakano. Deterministic annealing em algorithm. *Neural networks*, 11(2):271–282, 1998.

[97] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1667–1675, 2017.

[98] Rizwan Ahmed Khan, Alexandre Meyer, Hubert Konik, and Saida Bouakaz. Pain detection through shape and appearance features. In *2013 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2013.

[99] Reneiro Andal Virrey, Chandratilak De Silva Liyanage, Mohammad Iskandar bin Pg Hj Petra, and Pg Emeroylariffion Abas. Visual data of facial expressions for automatic pain detection. *Journal of Visual Communication and Image Representation*, 61:209–217, 2019.

[100] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE international conference on cybernetics (CYBCO)*, pages 128–131. IEEE, 2013.

[101] Ruijing Yang, Shujun Tong, Miguel Bordallo, Elhocine Boutellaa, Jinye Peng, Xiaoyi Feng, and Abdenour Hadid. On pain assessment from facial videos using spatio-temporal local

descriptors. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2016.

[102] N. Bouguila. Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198, 2011.

[103] Nuha Zamzami and Nizar Bouguila. A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognition*, 95:36–47, 2019.

[104] Pantea Koochemeshkian, Nuha Zamzami, and Nizar Bouguila. Flexible distribution-based regression models for count data: Application to medical diagnosis. *Cybernetics and Systems*, pages 1–25, 2020.

[105] Fatma Najar, Nuha Zamzami, and Nizar Bouguila. Fake news detection using bayesian inference. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 389–394. IEEE, 2019.

[106] Nizar Bouguila. A data-driven mixture kernel for count data classification using support vector machines. In *2008 IEEE Workshop on Machine Learning for Signal Processing*, pages 26–31. IEEE, 2008.

[107] Martin Ridout, Clarice GB Demétrio, and John Hinde. Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192. International Biometric Society Invited Papers Cape Town, South Africa, 1998.

[108] Wan Tang, Naiji Lu, Tian Chen, Wenjuan Wang, Douglas David Gunzler, Yu Han, and Xin M Tu. On performance of parametric and distribution-free models for zero-inflated and over-dispersed count responses. *Statistics in medicine*, 34(24):3235–3245, 2015.

[109] Tammy Harris, Joseph M Hilbe, and James W Hardin. Modeling count data with generalized distributions. *The Stata Journal*, 14(3):562–579, 2014.

[110] Joseph M. Hilbe. *Modeling Count Data*, pages 836–839. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[111] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2): 102–107, 2016.

[112] Minghui Huang, Haoran Xie, Yanghui Rao, Yuwei Liu, Leonard KM Poon, and Fu Lee Wang. Lexicon-based sentiment convolutional neural networks for online review analysis. *IEEE Transactions on Affective Computing*, pages 1–1, 2020.

[113] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*, pages 1–1, 2020.

[114] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[115] Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[116] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[117] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U Rajendra Acharya. Abcdm: An attention-based bidirectional cnn-rnn deep model for sentiment analysis. *Future Generation Computer Systems*, 115:279–294, 2021.

[118] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE, 2017.

[119] Md Shad Akhtar, Dushyant Singh Chauhan, and Asif Ekbal. A deep multi-task contextual attention framework for multi-modal affect analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(3):1–27, 2020.

[120] Navonil Majumder, Soujanya Poria, Haiyun Peng, Niyati Chhaya, Erik Cambria, and Alexander Gelbukh. Sentiment and sarcasm classification with multitask learning. *IEEE Intelligent Systems*, 34(3):38–43, 2019.

[121] Erik Cambria, Yang Li, Frank Z Xing, Soujanya Poria, and Kenneth Kwok. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 105–114, 2020.

[122] Ke Zhang, Yuanqing Li, Jingyu Wang, Erik Cambria, and Xuelong Li. Real-time video emotion recognition based on reinforcement learning and domain knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[123] Nizar Bouguila and Djemel Ziou. High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1716–1731, 2007.

[124] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons, 2011.

[125] Guo-Liang Tian, Kai Wang Ng, and Zhi Geng. Bayesian computation for contingency tables with incomplete cell-counts. *Statistica Sinica*, pages 189–206, 2003.

[126] Nuha Zamzami and Nizar Bouguila. Probabilistic modeling for frequency vectors using a flexible shifted-scaled dirichlet distribution prior. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(6):1–35, 2020.

[127] Manas Somaiya, Christopher Jermaine, and Sanjay Ranka. Learning correlations using the mixture-of-subsets model. *ACM Trans. Knowl. Discov. Data*, 1(4), February 2008.

[128] Xiao-Li Meng and David Van Dyk. The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3): 511–567, 1997.

[129] Zhanyu Ma, Yuping Lai, W Bastiaan Kleijn, Yi-Zhe Song, Liang Wang, and Jun Guo. Variational bayesian learning for dirichlet process mixture of inverted dirichlet distributions in non-gaussian image feature modeling. *IEEE transactions on neural networks and learning systems*, 30(2):449–463, 2018.

[130] Ma Zhanyu, Lai Yuping, Xie Jiyang, Meng Deyu, Kleijn W. Bastiaan, Guo Jun, and Yu Jingyi. Dirichlet process mixture of generalized inverted dirichlet distributions for positive vector data with extended variational inference. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

[131] Mithun Das Gupta, Srinidhi Srinivasa, Meryl Antony, et al. Kl divergence based agglomerative clustering for automated vitiligo grading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2700–2709, 2015.

[132] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *ICDM*, volume 4, pages 19–26, 2004.

[133] Bin Jiang, Jian Pei, Yufei Tao, and Xuemin Lin. Clustering uncertain data based on probability distribution similarity. *IEEE Transactions on Knowledge and Data Engineering*, 25(4): 751–763, 2011.

[134] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[135] Tony Jebara and Risi Kondor. Bhattacharyya and expected likelihood kernels. In *Learning theory and kernel machines*, pages 57–71. Springer, 2003.

[136] Bruce G Lindsay et al. Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The annals of statistics*, 22(2):1081–1114, 1994.

[137] Jacob Burbea and C Radhakrishna Rao. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *Journal of Multivariate Analysis*, 12(4): 575–596, 1982.

[138] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

[139] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[140] Frank Nielsen and Sylvain Boltz. The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57(8):5455–5466, 2011.

[141] Zhangcheng Qiu and Hong Shen. User clustering in a dynamic social network topic model for short text streams. *Information Sciences*, 414:102–116, 2017.

[142] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. KDD '14, page 233–242, New York, NY, USA, 2014. Association for Computing Machinery.

[143] Ian Holmes, Keith Harris, and Christopher Quince. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one*, 7(2):1–15, 2012.

[144] Franklin A Graybill. *Matrices with applications in statistics*. Number 512.896 G7 1983. 1983.

[145] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*, 2016.

[146] Conrad Tucker, Barton K Pursel, and Anna Divinsky. Mining student-generated textual data in moocs and quantifying their effects on student performance and learning outcomes. In *2014 ASEE Annual Conference & Exposition*, pages 24–907, 2014.

[147] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638, 2014.

[148] Peiying Zhang, Xingzhe Huang, and Lei Zhang. Information mining and similarity computation for semi-/un-structured sentences from the social data. *Digital Communications and Networks*, 2020.

[149] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

[150] Ebba Cecilia Ovesdotter Alm. *Affect in text and speech*. University of Illinois at Urbana-Champaign, 2008.

[151] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer, 2015.

[152] Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, 2010.

[153] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[154] Rana el Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[155] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International conference on pattern recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE, 2006.

[156] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008.

[157] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1):329–349, 2003.

[158] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[159] Lin Qiu, Yong Cao, Zaiqing Nie, Yong Yu, and Yong Rui. Learning word representation considering proximity and ambiguity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[160] Xinzhi Wang, Hui Zhang, and Yi Liu. Sentence vector model based on implicit word vector expression. *IEEE Access*, 6:17455–17463, 2018.

[161] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[162] C Lee Giles, Gary M Kuhn, and Ronald J Williams. Dynamic recurrent neural networks: Theory and applications. *IEEE Transactions on Neural Networks*, 5(2):153–156, 1994.

[163] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

[164] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[165] Saif M Mohammad. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741, 2012.

[166] Ekaterina P Volkova, Betty Mohler, Detmar Meurers, Dale Gerdemann, and Heinrich H Bülthoff. Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106, 2010.

[167] Alberto Acerbi, Vasileios Lampos, and R Alexander Bentley. Robustness of emotion extraction from 20 th century english books. In *2013 IEEE International Conference on Big Data*, pages 1–8. IEEE, 2013.

[168] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. Mining social emotions from affective text. *IEEE transactions on knowledge and data engineering*, 24(9):1658–1670, 2011.

[169] Jianhui Pang, Yanghui Rao, Haoran Xie, Xizhao Wang, Fu Lee Wang, Tak-Lam Wong, and Qing Li. Fast supervised topic models for short text emotion detection. *IEEE Transactions on Cybernetics*, 51(2):815–828, 2019.

[170] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433, 2006.

[171] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180, 2007.

[172] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1424–1445, 2000.

[173] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2012.

[174] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio

Escalera, and Gholamreza Anbarjafari. Survey on emotional body gesture recognition. *IEEE transactions on affective computing*, 12(2):505–523, 2018.

[175] Hatice Gunes and Massimo Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):64–84, 2008.

[176] Tadas Baltrušaitis, Daniel McDuff, Ntombikayise Banda, Marwa Mahmoud, Rana El Kaliouby, Peter Robinson, and Rosalind Picard. Real-time inference of mental states from facial expressions and upper body gestures. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 909–914. IEEE, 2011.

[177] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, et al. The humaine database. In *Emotion-Oriented Systems*, pages 243–284. Springer, 2011.

[178] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[179] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1):177–196, 2001.

[180] Alexander Hinneburg, Hans-Henning Gabriel, and Andre Gohr. Bayesian folding-in with dirichlet kernels for plsi. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 499–504. IEEE, 2007.

[181] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[182] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[183] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[184] Peter V Gehler, Alex D Holub, and Max Welling. The rate adapting poisson model for information retrieval and object recognition. In *Proceedings of the 23rd international conference on Machine learning*, pages 337–344, 2006.

[185] Thomas Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in neural information processing systems*, pages 914–920, 2000.

[186] Cecilia Ovesdotter Alm and Richard Sproat. Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction*, pages 668–674. Springer, 2005.

[187] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[188] Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *European conference on information retrieval*, pages 29–41. Springer, 2009.

[189] Nizar Bouguila and Djerriel Ziou. Improving content based image retrieval systems using finite multinomial dirichlet mixture. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pages 23–32. IEEE, 2004.

[190] Shao Lei Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2. IEEE, 2004.

[191] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[192] Toni M Rath, Victor Lavrenko, and R Manmatha. A statistical approach to retrieving historical manuscript images without recognition. Technical report, Space and Naval Warfare Systems Center San Diego CA, 2003.

[193] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 515–524, 2017.

[194] Anderson Uilian Kauer and Viviane P Moreira. Using information retrieval for sentiment polarity prediction. *Expert Systems with Applications*, 61:282–289, 2016.

[195] Amal Htait, Sébastien Fournier, Patrice Bellot, Leif Azzopardi, and Gabriella Pasi. Using sentiment analysis for pseudo-relevance feedback in social book search. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 29–32, 2020.

[196] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42th annual meeting of the association of computational linguistics (ACL)*, pages 271–278, 2004.

[197] Soroush Vosoughi, Helen Zhou, and Deb Roy. Enhanced twitter sentiment classification using contextual information. *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2016.

[198] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, 2010.

[199] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, pages 655–666, 2014.

[200] Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, 2014.

[201] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260, 2018.

[202] Giulio Petrucci and Mauro Dragoni. An information retrieval-based system for multi-domain sentiment analysis. In *Semantic Web Evaluation Challenges*, pages 234–243. Springer, 2015.

[203] Victor Lavrenko and W Bruce Croft. Relevance-based language models. In *ACM SIGIR Forum*, volume 51, pages 260–267. ACM New York, NY, USA, 2017.

[204] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

[205] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

[206] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 55–64, 2016.

[207] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 375–384, 2018.

[208] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard De Melo. Co-pacrr: A context-aware neural ir model for ad-hoc retrieval. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 279–287, 2018.

[209] Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[210] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

[211] Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. Modeling interestingness with deep neural networks. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[212] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web*, pages 373–374, 2014.

[213] Stephen E Robertson. The probability ranking principle in ir. *Journal of documentation*, 1977.

[214] Donald Metzler, Victor Lavrenko, and W Bruce Croft. Formal multiple-bernoulli models for language modeling. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 540–541, 2004.

[215] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.

[216] Bhaskar Mitra and Nick Craswell. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*, 2017.

[217] Yufeng Zhang, Jinghao Zhang, Zeyu Cui, Shu Wu, and Liang Wang. A graph-based relevance matching model for ad-hoc retrieval. *arXiv preprint arXiv:2101.11873*, 2021.

[218] Gianna Serafina Monti, Gloria Mateu-Figueras, and Vera Pawlowsky-Glahn. Notes on the scaled dirichlet distribution. *Compositional data analysis*, pages 128–138, 2011.

[219] Victor Lavrenko. *A generative theory of relevance*. PhD thesis, 2004.

[220] Norbert Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[221] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[222] H. Saif, M. Fernandez, Y. He, and H. Alani. Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold. *CEUR Workshop Proceedings*, 1096:9–21, 2013.

[223] Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. SemEval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[224] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1): 163–173, 2012.

[225] Luiz FS Coletta, Nadia FF da Silva, Eduardo Raul Hruschka, and Estevam Rafael Hruschka. Combining classification and clustering for tweet sentiment analysis. In *2014 Brazilian conference on intelligent systems*, pages 210–215. IEEE, 2014.

[226] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[227] Joseph M Hilbe. *Modeling count data*. Cambridge University Press, 2014.

[228] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[229] David M Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*, volume 24, pages 411–418. Citeseer, 2008.

[230] Alessio Benavoli and Cassio P De Campos. Inference from multinomial data based on a mle-dominance criterion. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 22–33. Springer, 2009.

[231] Cassio Polpo de Campos and Alessio Benavoli. Inference with multinomial data: Why to weaken the prior strength. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[232] Alan Agresti and David B Hitchcock. Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14(3):297–330, 2005.

[233] Balaji Krishnapuram, Lawrence Carin, Mário AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005.

[234] Marcus Hutter. Sparse adaptive dirichlet-multinomial-like processes. In *Conference on Learning Theory*, pages 432–459. PMLR, 2013.

[235] Eric Sven Ristad. A natural law of succession. Technical report, Department of Computer Science, Princeton University, July 1998.

[236] Thomas L. Griffiths and Joshua B. Tenenbaum. Using vocabulary knowledge in bayesian multinomial estimation. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 1385–1392, Cambridge, MA, USA, 2001. MIT Press.

[237] Nir Friedman and Yoram Singer. Efficient bayesian parameter estimation in large discrete domains. In *Advances in neural information processing systems*, pages 417–423, 1999.

[238] William Barcella, Maria De Iorio, Stefano Favaro, and Gary L Rosner. Dependent generalized dirichlet process priors for the analysis of acute lymphoblastic leukemia. *Biostatistics*, 19(3):342–358, 2018.

[239] Elise Epaillard and Nizar Bouguila. Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection. *IEEE transactions on neural networks and learning systems*, 30(4):1034–1047, 2018.

[240] Susana Eyheramendy, David D Lewis, and David Madigan. On the naive bayes model for text categorization. 2003.

[241] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[242] Nizar Bouguila. A model-based approach for discrete data clustering and feature weighting using MAP and stochastic complexity. *IEEE Trans. Knowl. Data Eng.*, 21(12):1649–1664, 2009.

[243] Muhammad Abbas, K Ali Memon, A Aleem Jamali, Saleemullah Memon, and Anees Ahmed. Multinomial naive bayes classification model for sentiment analysis. *IJCSNS*, 19 (3):62, 2019.

[244] Sumedh Kadam, Aayush Gala, Pritesh Gehlot, Aditya Kurup, and Kranti Ghag. Word embedding based multinomial naive bayes algorithm for spam filtering. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–5. IEEE, 2018.

[245] Quan Yuan, Gao Cong, and Nadia Magnenat Thalmann. Enhancing naive bayes with various smoothing methods for short text classification. In *Proceedings of the 21st International Conference on World Wide Web*, pages 645–646, 2012.

[246] Don Willems and Louis Vuurpijl. A bayesian network approach to mode detection for interactive maps. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 869–873. IEEE, 2007.

[247] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[248] Jing Bai, Jian-Yun Nie, and François Paradis. Using language models for text classification. In *Proceedings of the Asia Information Retrieval Symposium, Beijing, China*, 2004.

[249] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2): 179–214, 2004.

[250] Jian Zhang, Zoubin Ghahramani, and Yiming Yang. A probabilistic model for online document clustering with application to novelty detection. *Advances in neural information processing systems*, 17:1617–1624, 2004.

[251] Nizar Bouguila and Mukti Nath Ghimire. Discrete visual features modeling via leave-one-out likelihood estimation and applications. *J. Vis. Commun. Image Represent.*, 21(7):613–626, 2010.

[252] Nizar Bouguila. On the smoothing of multinomial estimates using liouville mixture models and applications. *Pattern Anal. Appl.*, 16(3):349–363, 2013.

[253] Nir Friedman Yoram Singer. Efficient bayesian parameter estimation in large discrete domains. *Advances in neural information processing systems*, 11:417, 1999.

[254] Elise Epaillard and Nizar Bouguila. Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognit.*, 55:125–136, 2016.

[255] Nizar Bouguila. Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognit. Lett.*, 33(2):103–110, 2012.

[256] Wentao Fan and Nizar Bouguila. Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE Trans. Neural Networks Learn. Syst.*, 24(11):1850–1862, 2013.

[257] Wentao Fan and Nizar Bouguila. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1323–1329. IJCAI/AAAI, 2013.

[258] Saif Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, aug 2017. Association for Computational Linguistics.

[259] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

[260] T HOFMANN. Probabilistic latent semantic analysis. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI), 1999*, pages 289–296, 1999.

[261] Martin Riedl and Chris Biemann. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 student research workshop*, pages 37–42, 2012.

[262] Ramnath Balasubramanyan and William W Cohen. Block-lda: Jointly modeling entity-annotated text and entity-entity links. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 450–461. SIAM, 2011.

[263] Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, 163:1–13, 2019.

[264] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048, 2011.

[265] Qi Liu, Enhong Chen, Hui Xiong, Chris HQ Ding, and Jian Chen. Enhancing collaborative filtering by user interest expansion via personalized ranking. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):218–233, 2011.

[266] Chong Wang and David Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. *Advances in neural information processing systems*, 22:1982–1989, 2009.

[267] Sinead Williamson, Chong Wang, Katherine Heller, and David Blei. Focused topic models. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, pages 1–4, 2009.

[268] Tianyi Lin, Wentao Tian, Qiaozhu Mei, and Hong Cheng. The dual-sparse topic model: mining focused topics and focused terms in short text. In *Proceedings of the 23rd international conference on World wide web*, pages 539–550, 2014.

[269] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[270] Cheng Zhang, Carl Ek, Xavi Gratal, Florian Pokorny, and Hedvig Kjellstrom. Supervised hierarchical dirichlet processes with variational inference. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 254–261, 2013.

[271] Oksana Yakhnenko and Vasant Honavar. Multi-modal hierarchical dirichlet process model for predicting image annotation and image-object label correspondence. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 283–293. SIAM, 2009.

[272] Thomas L Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(4), 2011.

[273] Sinead Williamson, Chong Wang, Katherine A Heller, and David M Blei. The ibp compound dirichlet process and its application to focused topic modeling. In *ICML*, 2010.

[274] François Caron, Manuel Davy, and Arnaud Doucet. Generalized polya urn for time-varying dirichlet process mixtures. UAI'07, page 33–40, Arlington, Virginia, USA, 2007. AUAI Press. ISBN 0974903930.

[275] Jason A Duan, Michele Guindani, and Alan E Gelfand. Generalized spatial dirichlet process models. *Biometrika*, 94(4):809–825, 2007.

[276] Jim E Griffin and MF J Steel. Order-based dependent dirichlet processes. *Journal of the American statistical Association*, 101(473):179–194, 2006.

[277] Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani. Dependent indian buffet processes. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 924–931. JMLR Workshop and Conference Proceedings, 2010.

[278] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(8), 2011.

[279] Samuel J Gershman, Peter I Frazier, and David M Blei. Distance dependent infinite latent feature models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):334–345, 2014.

[280] Yezheng Liu, Fei Du, Jianshan Sun, and Yuanchun Jiang. ilda: An interactive latent dirichlet allocation model to improve topic quality. *Journal of Information Science*, 46(1):23–40, 2020.

[281] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[282] Aaron C Courville, James Bergstra, and Yoshua Bengio. Unsupervised models of images by spikeand-slab rbms. In *ICML*, 2011.

[283] Aaron Courville, James Bergstra, and Yoshua Bengio. A spike and slab restricted boltzmann machine. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 233–241. JMLR Workshop and Conference Proceedings, 2011.

[284] Ali Faisal, Jussi Gillberg, Gayle Leen, and Jaakko Peltonen. Transfer learning using a non-parametric sparse topic model. *Neurocomputing*, 112:124–137, 2013.

[285] Bingshan Zhu, Yi Cai, and Huakui Zhang. Sparse biterm topic model for short texts. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 227–241. Springer, 2021.

[286] Jun Zhu and Eric P Xing. Sparse topical coding. In *Proceedings of the 27th international conference on uncertainty in artificial intelligence*, 2012.

[287] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1):157–208, 2012.

[288] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. Short and sparse text topic modeling via self-aggregation. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[289] Qianqian Xie, Prayag Tiwari, Deepak Gupta, Jimin Huang, and Min Peng. Neural variational sparse topic model for sparse explainable text representation. *Information Processing & Management*, 58(5):102614, 2021.

[290] Hyun Duk Kim and ChengXiang Zhai. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 385–394, 2009.

[291] Yuanbin Wu, Qi Zhang, Xuan-Jing Huang, and Lide Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1533–1541, 2009.

[292] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9(1):1–20, 2019.

[293] Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, 2015.

[294] Aytug Onan, Serdar Korukoglu, and Hasan Bulut. Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7(1):101–119, 2016.