Transformer-Based Models for Identifying Customer Needs in User-Generated Content: Performance Gaps, Unintended Bias, and Broader Implications

Mehrshad Kashi

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of
Master of Applied Science (Electrical and Computer Engineering) at
Concordia University
Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Mehrshad Kashi**
Entitled: **Transformer-Based Models for Identifying Customer Needs in User-Generated Content: Performance Gaps, Unintended Bias, and Broader Implications**

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

Master of Applied Science (Electrical and Computer 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. Sébastien Le Beux

_____ Examiner
Dr. Dongliang Sheng

_____ Supervisor
Dr. Otmane Ait Mohamed

_____ Co-supervisor
Dr. Salim Lahmiri

Approved by _____
Dr. Abdelwahab Hamou-Lhadj, Chair
Department of Electrical and Computer Engineering

_____ 2025

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Transformer-Based Models for Identifying Customer Needs in User-Generated Content: Performance Gaps, Unintended Bias, and Broader Implications

Mehrshad Kashi

This thesis reviews and evaluates intelligent methods for identifying customer needs in user-generated content (UGC). It first surveys prior work and shows that many studies share generic goals yet overlook the complexity and taxonomy of needs in their evaluation setups. To clarify scope, the thesis distinguishes between using Machine Learning (ML) as a tool to support marketing workflows and treating customer-needs identification itself as an Natural Language Processing (NLP) task with clear definitions and constructs. Building on this perspective, a large experimental study assesses Transformer-based models for generalizability, robustness, fairness, and sample efficiency across varied settings. Results indicate competitive accuracy, with gains in F1 up to 18% over baselines, but also consistent limitations: shared error patterns, difficulty with rare or unseen needs, reliance on lexical cues that weakens cross-domain performance, and no guaranteed gains in sample efficiency from larger models. Cross-domain results benefit most from richer, diverse domain training, while adding more in-domain data does not improve transfer. Beyond technical metrics, the thesis highlights adoption barriers, costs, data constraints, task complexity, and ethical considerations and argues for evaluation frameworks that reflect taxonomy, transparency, and fairness. It concludes with practical guidance that bridges marketing theory and NLP practice to support responsible, reproducible deployment.

Acknowledgments

I would like to sincerely thank Dr. Otmane Ait Mohamed, my supervisor, for giving me the chance to be part of the HVG lab. Working under his guidance has been an invaluable experience, and I truly appreciate his constant support, patience, and encouragement throughout this journey. The HVG lab was not just a place to work, it was a friendly and welcoming environment, like a family, and I truly enjoyed being there.

I would also like to thank Dr. Salim Lahmiri, my co-supervisor, for his guidance and valuable feedback. His advice has always pushed me to look at my work more carefully and to improve it in meaningful ways.

My deepest gratitude goes to my parents, Leila and Majid. They have always supported me in every possible way, without conditions and without hesitation. I would not have been able to reach this stage without their endless encouragement and belief in me.

I am also deeply thankful to my partner, Shirin, who has never stopped believing in me. She has been by my side at every single step of this journey, and especially in the final stage, when she gave me all her time and energy. I cannot find words strong enough to thank her.

Finally, I want to thank my brother, Mahyar. Even though distance has kept us apart for many years, his support and love have always been with me. I miss him more than anyone else, and I dedicate this small piece of work to him.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. 5, 13, 30

CN customer needs. 20–22, 24–26, 30, 32, 33

DL Deep Learning. 4, 5, 8

ISCN Identifying Sentences containing Customer Needs. 2–4, 18, 23, 65, 66

LLM Large Language Model. 5, 11–15, 18, 19, 24, 26, 32, 33, 36

ML Machine Learning. iii, 1, 3, 10, 12, 13, 18, 20–25, 27, 29, 32, 33, 39, 65

NLP Natural Language Processing. iii, 1, 4–8, 10, 12, 17, 20–22, 25, 26, 32, 33, 36, 65

PSA Perturbation Sensitivity Analysis. 14

RE requirements engineering. 21, 34

SOTA state-of-the-art. 2–4, 26

UGC user-generated content. iii, 1–5, 17–22, 24–27, 29, 30, 32, 33, 36, 65

Chapter 1

Thesis Introduction

1.1 Introduction

In today's digital era, vast amounts of UGC are shared online every day, creating significant opportunities to extract meaningful insights. Exploiting these opportunities requires developing efficient and scalable models capable of analyzing large volumes of data. Such analytical approaches have diverse applications, ranging from monitoring consumer sentiment to improve customer satisfaction (Gonzalez, 2019), and forecasting stock market trends (Bollen, Mao, & Zeng, 2011), to early detection of depressive symptoms and assessment of online users' suicidal risk (De Choudhury, Gamon, Counts, & Horvitz, 2021; O'Dea et al., 2015).

Customer reviews are a unique type of UGC (often in textual format) that reflect personal experiences and expectations with a product. Unlike quantitative metrics, these narratives offer richer and more nuanced insights, often uncovering subtle patterns in how consumers experience products and perceive brands. Moreover, the public availability of customer reviews helps build trust and transparency, influencing potential buyers and reinforcing a brand's credibility. An emerging application of such qualitative feedback is the extraction of customer needs (Kuehl, Scheurenbrand, & Satzger, 2016; Timoshenko & Hauser, 2019), which enables organizations to pinpoint specific areas for product improvement (Guo et al., 2016), refine their product ecosystems (Zhou, Ayoub, Xu, & Jessie Yang, 2020), and adjust marketing strategies in response to emerging consumer trends (D. T. S. Kumar, 2020).

However, extracting customer needs from the vast volume of online reviews presents several challenges. Manual analysis is slow, inconsistent, and expensive, making it impractical for real-time or large-scale applications. To address this, organizations can adopt intelligent systems to process customer feedback in fully or semi-automated workflows. By leveraging ML-based techniques, these systems create scalable and continuous feedback loops that convert qualitative input into structured insights, enabling ongoing innovation and supporting strategic decision-making while minimizing operational costs (Kuehl, Mühlthaler, & Goutier, 2020).

Despite recent advances in NLP, particularly the development of Transformer-based models (Vaswani et al., 2017), that have substantially improved the process of customer needs identification from UGC, several important limitations continue to hinder their broader effectiveness. Among these, generalization is one of the main challenges, as in low-resource settings, insufficient data can severely limit model robustness and predictive accuracy. In addition, Transformer models often inherit and amplify existing societal biases embedded in training data, raising ethical and practical concerns regarding fairness and representation in decision-making. While the presence of human

oversight can partially compensate for some of these issues (e.g., poor generalization and biased predictions), the field is rapidly advancing toward fully automated agentic systems capable of managing the entire pipeline—from customer feedback analysis to product adaptation (C. Wang, Jiang, Li, Hu, & Lin, 2024)—making it essential to examine these limitations and mitigate the risks they pose (Dhamodharan, 2025).

Although recent research has shown increasing interest in developing intelligent methods for customer needs detection, most efforts have focused narrowly on improving predictive accuracy, with far less attention given to designing comprehensive evaluation frameworks that consider not only technical performance but also crucial real-world concerns such as robustness in low-resource settings and the amplification of societal biases, which directly impact deployment in practice.

Building on the limitations identified in recent work, the main objective of this thesis is threefold: (1) to identify and analyze the specific challenges that constrain the performance of state-of-the-art (SOTA) models in the task of identifying customer needs from UGC; (2) to conduct a comparative analysis of current SOTA methods (i.e., Transformer-based models) highlighting their limitations in domain-specific contexts; and (3) to investigate whether unintended social biases, emerging during model pretraining or development, affect model decisions in ways that may lead to discriminatory or neglectful outcomes. By shedding light on underexplored challenges through a critical review of the literature, and by comprehensively evaluating existing methods from multiple perspectives, this study aims to support the development of approaches that improve model performance while mitigating algorithmic biases and minimizing the risk of social harm.

1.2 Problem Statement

In customer needs detection from UGC, existing methods typically follow one of two strategies: predicting the exact category of the expressed need (e.g., through multi-class classification), effectively framing the task as identifying *"what"* need is present in the context (Kuehl et al., 2016); or predicting whether the context contains any customer need at all, framing the task as determining *"whether"* a need is present (Timoshenko & Hauser, 2019). The former relies on a predefined taxonomy of customer need types and assumes these needs are fixed and well-defined, an assumption that often fails to hold in real-world settings such as UGC, where needs are dynamic, nuanced, and require human interpretation. In addition, narrowly fixing the label space can limit the detection of rare or novel needs, which often reflect unmet demands and can offer companies valuable insights for product innovation (Kärkkäinen, Piippo, Puumalainen, & Tuominen, 2001; von Hippel, 1986).

In contrast, a binary classification approach, which labels each context as either "informative" (i.e., containing a customer need) or "non-informative," is often more straightforward and offers a more flexible and practical alternative. While this formulation does not specify the type of need, it serves as an effective filtering stage to reduce the volume of UGC that must be reviewed by professionals and tends to be more robust in dynamic settings where unseen or uncommon needs may emerge. Such binary text classification formulation has also been shown to be cost-effective from an operational standpoint. As reported by Timoshenko and Hauser (2019), this filtering approach can reduce the time required by marketing professionals to extract detailed customer needs by approximately 45–55%.

Based on these considerations, this thesis focuses on identifying customer needs as expressed in sentence-level contexts within UGC, framing the task as determining whether a given sentence contains a customer need. Throughout this thesis, we refer to this binary classification task as Identifying Sentences containing Customer Needs (ISCN) and utilize online reviews as the data

source.

1.3 Research Gaps

As businesses increasingly rely on UGC to identify customer needs, the focus has shifted from justifying the use of ML solutions to addressing the technical and societal barriers that hinder their robust adoption. These barriers include high economic costs of ML research and deployment, technical challenges in data availability and model efficacy, and social considerations such as discrimination and environmental impact (Cubric, 2020; D. Kumar & Suthar, 2024).

No research has specifically addressed the challenges that constrain the performance of SOTA models in customer needs identification tasks, particularly in the ISCN. Although prior studies underscore the importance of scalable and reliable models, the performance limitations of Transformer-based architectures remain underexamined. Moreover, no comprehensive evaluation framework has been proposed to thoroughly assess the societal implications of intelligent methods in this domain, an essential step toward ensuring responsible AI adoption.

1.4 Thesis Objectives

Considering the research gaps outlined in Section 1.3, the primary objective of this research is not to develop a novel approach or to enhance the performance of state-of-the-art (SOTA) models in the ISCN task. Rather, this study examines the key technical that must be considered when designing a practical and effective needs identification system. To guide the comprehensive analysis presented throughout this thesis, the following research questions are proposed:

- **RQ1: What challenges and domain complexities characterize the ISCN task, how do they affect model robustness, and why should they be incorporated into evaluation frameworks?**

  Existing literature on the ISCN task predominantly focuses on developing new or improved model architectures while often overlooking fundamental challenges that can distort performance and hinder the deployment of robust real-world solutions. This research question aims to deepen our understanding of these issues, thereby encouraging the adoption of more nuanced evaluation strategies and enhanced data practices to sustain model performance in evolving UGC scenarios.

- **RQ2: What comprehensive evaluation framework can be developed to determine whether an ML model for customer needs analysis is both sufficiently effective and superior to alternative approaches?**

  This inquiry addresses the dual challenge of evaluating whether an ML model for customer needs analysis meets baseline performance standards while also outperforming alternative approaches under task-specific criteria. It seeks to establish an evaluation framework that extends beyond traditional accuracy metrics by incorporating dimensions such as generalization, robustness, fairness, and adaptability. In particular, the inquiry examines how models handle complex, dynamic data and address challenges like limited annotated datasets, sample selection bias, and evolving data trends. The anticipated outcome is to guide the selection and adoption of models that offer both operational effectiveness and a competitive advantage in the rapidly evolving landscape of customer needs analysis.

1.5 Thesis Contributions

This thesis makes the following contributions:

- We provided a comprehensive discussion of the overlooked challenges that affect real-world performance in the ISCN classification task, including annotation inconsistencies, selection bias, temporal shifts, and linguistic complexities. By consolidating these issues through a critical review of literature, we underscore the importance of moving beyond simple metrics and integrating such considerations into subsequent evaluation frameworks for more robust and adaptable model design (see Chapter 3). A structured review conducted in this section further supports this argument.
- We Proposed an evaluation framework for the ISCN classification task, moving beyond conventional metrics to assess model performance from both quality and quantitative perspectives under conditions such as domain shifts and low-frequency needs with respect to discussed challenges in Chapter 3. This framework highlights the importance of generalization, robustness, fairness, and adaptability in real-world scenarios (see Chapter 4).

1.6 Structure of the Thesis

This chapter serves as an introduction to the identification of customer needs from UGC and highlights the motivation for applying SOTA techniques to address challenges in analyzing UGC. Chapter 2 presents an overview of relevant concepts in NLP, Deep Learning (DL), and the issues of bias and fairness in NLP systems, and concludes with a review of key literature on customer needs analysis. Chapter 3 discusses the various complexities and challenges associated with identifying customer needs in UGC and reviews recent literature. Chapter 4 provides a comprehensive analysis of Transformer-based models by investigating their technical limitations in the context of ISCN and guiding future research aimed at overcoming these barriers for real-world adoption. Finally, Chapter 5 concludes the thesis by summarizing the principal findings, discussing the limitations of the current research, and proposing several directions for future work.

Chapter 2

Background

This chapter reviews foundational and recent literature that supports the core contributions of this thesis. It also includes a forward-looking discussion on how recent advancements in generative language models can further enhance this field.

Section 2.1 introduces the fundamentals of NLP by discussing classical text representation methods, foundational word embedding techniques, and the evaluation metrics employed throughout this thesis. Section 2.2 shifts focus to DL for language modeling, with particular attention to the emergence of the Transformer architecture and the subsequent development of Transformer-based language models. Section 2.3 provides preliminaries for bias and fairness in NLP systems, detailing relevant terminology and taxonomy, potential implications, and strategies for both evaluation and mitigation, while also highlighting underlying sources and root causes. In Section 2.4, a comprehensive review of existing literature on customer needs analysis is presented, encompassing marketing perspectives, the role of UGC in needs elicitation, and intelligent methods for effectively identifying customer requirements from these data sources. Finally, Section 2.5 explores how generative Large Language Models (LLMs) can be leveraged to fully automate the process of eliciting customer needs, thereby offering insights into more effective and data-driven approaches.

2.1 Natural language Processing

NLP (also known as computational linguistics) is a multidisciplinary field at the intersection of linguistics and computer science that aims to enable computers to understand, interpret, and generate human language (Manning & Schütze, 1999) by integrating computational methods with techniques from various subfields of Artificial Intelligence (AI), including machine learning and deep learning (H. Li, 2017). Over time, NLP has evolved from early rule-based and statistical approaches to more advanced machine learning-driven techniques. Initial efforts focused on fundamental tasks such as part-of-speech tagging, named entity recognition, and syntactic and semantic parsing, which laid the foundation for early applications such as machine translation and document summarization. The advent of large-scale neural architectures, along with advancements in hardware, has further enhanced the field's capacity for nuanced language understanding and generation. These developments have driven broader adoption in real-world applications and transformed human-computer interaction through sophisticated conversational agents across diverse domains.

Closely related to NLP, text mining is the process of extracting meaningful information, patterns, and insights from unstructured textual data (Feldman & Dagan, 1995) through data-driven

techniques from statistical analysis and NLP. Developing such systematic pattern recognition systems is highly beneficial across various industries, facilitating informed decision-making and enhancing analytical capabilities (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). For example, in healthcare, analyzing clinical notes and medical records can reveal critical trends or aid in diagnosis of fatal conditions such as cancer (Sheikhalishahi et al., 2019). In finance, automated systems use text mining to assess market sentiment, flag risks, and predict stock price movements (Du, Xing, Mao, & Cambria, 2024). In marketing research, firms can analyze the voice of the customer to align their strategies with evolving market demands, preferences, and expectations (Griffin & Hauser, 1993).

As the volume of unstructured text data continues to grow exponentially—through social media, news articles, customer feedback, and other digital content—the integration of advanced NLP and text mining techniques has become increasingly crucial for businesses. These technologies not only drive automation but also enhance the accuracy and depth of insights derived from large-scale textual data.

2.1.1 Classical Approaches for Text Representation

Unlike humans, computers require structured representations to process text. Text representation techniques transform unstructured text into structured numerical formats, making them suitable for algorithmic processing. A common approach involves representing text as numerical vectors which facilitates computational analysis and building machine learning applications.

Traditional methods for text vectorization, such as the bag-of-words (BoW) model and term frequency-inverse document frequency (TF-IDF), are widely employed. The BoW model treats text as an unordered collection of words, disregarding grammatical structure, while TF-IDF assigns weights based on word frequency across documents, thereby emphasizing the significance of terms that are less common in the overall corpus. Despite their effectiveness in certain applications, these models often struggle to capture semantic relationships and are susceptible to issues such as high dimensionality and sparsity in word representations.

Regardless of their limitations, classical text vectorization methods remain foundational in certain NLP applications. While advancements in deep learning and contextual word embeddings have largely superseded traditional approaches, BoW and TF-IDF continue to be relevant in specific contexts. Their efficiency, interpretability, and scalability make them valuable in resource-constrained environments and applications where keyword presence is more critical than contextual meaning, such as information retrieval and search engine ranking algorithms.

2.1.2 Foundations of Continuous Word Representations

The concept of representing words as continuous vectors dates back several decades (Hinton, 1986). A significant breakthrough in this domain came with the development of Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013), which demonstrated the advantages of training word embeddings on large datasets using simple neural architectures. This approach introduced two shallow neural networks, Continuous Bag-of-Words (CBOW) and Skip-Gram, designed to compute continuous vector representations of words from extensive corpora. The CBOW model predicts a target word based on its surrounding context, whereas Skip-Gram performs the inverse operation by predicting surrounding words given a target word. These models effectively capture semantic relationships, enabling vector arithmetic for analogies (e.g., $v(\text{Beijing}) - v(\text{China}) + v(\text{France}) \approx v(\text{Paris})$). However, despite its success, Word2Vec primarily relies on local co-occurrence patterns, which limits its ability to represent out-of-vocabulary words and distinguish between different meanings of

polysemous words (e.g., "bank" as a financial institution vs. "bank" of a river), as it assigns a single static vector to each word regardless of context.

To overcome the limitations of Word2Vec, GloVe (Pennington, Socher, & Manning, 2014) introduced a global matrix factorization approach that leverages word co-occurrence statistics across an entire corpus. Unlike Word2Vec, which learns embeddings based on local context windows, GloVe constructs word representations by factorizing a word co-occurrence matrix and optimizing a weighted least squares objective. This approach enhances the encoding of global word associations, leading to improved representations of rare words and domain-specific vocabulary by capturing statistical relationships that may not be evident in local context windows. Although GloVe addressed some of Word2Vec's shortcomings by incorporating global corpus information, it still considers words as single units, preventing it from generalizing to unseen words and distinguishing multiple meanings of polysemous words.

Expanding the principles of Word2Vec and GloVe, FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) introduced subword embeddings, representing words as sequences of character n-grams. This approach captures morphological variations, allowing for more effective representations of rare and out-of-vocabulary words. By encoding subword information, FastText mitigates a key limitation of previous models, particularly in morphologically rich languages and low-resource scenarios.

While these word embedding techniques have significantly advanced NLP through their dense, interpretable representations, they remain inherently static—assigning a single vector to a word regardless of its contextual meaning. Recent advancements, particularly Transformer-based architectures (see Section 2.2), have addressed this challenge by generating context-dependent embeddings, enabling more nuanced language understanding and further advancing the field.

2.1.3 Evaluation Metrics

The performance of NLP models is multifaceted, encompassing dimensions such as predictive accuracy, robustness, fairness, and adaptability across diverse contexts. A comprehensive evaluation not only assesses how well a model generalizes to unseen data but also examines its consistency across varying conditions and its susceptibility to biases. The selection of appropriate evaluation metrics is crucial to ensuring reliable, fair, and effective deployment in real-world applications. While certain aspects of model performance, such as the identification of unintended biases, require specialized evaluation techniques (see Section 2.3.3), this section focuses on conventional classification-based metrics relevant to evaluating the performance of the classifier that serves as the core of a semi-automated system for customer needs identification within the ISCN framework.

Empirical evaluation measures to assess a classifier performance can be categorized into three groups: (1) metrics that offer qualitative insights into errors (e.g., accuracy, F1-score, precision, recall, and specificity), (2) metrics that capture a probabilistic view of errors (e.g., mean absolute error *(MAE)* and mean squared error *(MSE)*), and (3) metrics that evaluate how well the model ranks instances (e.g., AUC) (Ferri, Hernández-Orallo, & Modroiu, 2009). Each evaluation metric captures a distinct aspect of classification performance. For instance, high accuracy does not necessarily imply a high AUC. Accuracy reflects the overall proportion of correct classifications based on a fixed decision threshold, making it highly sensitive to class imbalance and threshold selection, while AUC evaluates the model's ability to distinguish between positive and negative instances across all possible thresholds, making it a more robust indicator of ranking performance. As a result, a model can achieve high accuracy while failing to separate classes, particularly in imbalanced datasets.

Therefore, relying on multiple measures provides a more comprehensive understanding of model capabilities, making them essential for robust evaluation.

The mathematical definitions of the classification metrics used in this study are provided in Section 4.4.2.

2.2 Deep Learning for Language Modeling

Language modeling has been a cornerstone of computational linguistics for several decades, with early explorations of neural networks in this domain found in the work of (Miikkulainen & Dyer, 1991). The pivotal breakthrough began with the introduction of neural probabilistic language models capable of learning distributed word representations (Bengio, Ducharme, Vincent, & Janvin, 2003). This approach significantly improved over traditional n-gram models, which suffer from data sparsity and lack generalization across syntactic and semantic contexts. Beyond improving word sequence prediction, it also reshaped perspectives on language modeling, demonstrating that neural networks, when trained on sufficiently large datasets, could capture linguistic structures more effectively than purely statistical methods.

Since then, the role of scaling in DL has become increasingly central to advancements in language modeling. Early studies established that hardware accelerators significantly enhance the efficiency of neural network training (Raina, Madhavan, & Ng, 2009), while subsequent work identified a strong correlation between model size and performance (Coates, Ng, & Lee, 2011). Later empirical findings further demonstrated that increasing both model size and training data leads to predictable performance gains, following a log-log scaling relationship (Hestness et al., 2017). These insights laid the foundation for a systematic recipe—leveraging large-scale and improved architectures, utilizing extensive datasets, and scaling computational resources- which continues to expand the frontiers of language modeling. An influential facilitator of this paradigm is the Transformer architecture, which has played a pivotal role in enabling large-scale training by introducing key innovations that enhance scaling efficiency and fundamentally reshape the field of NLP.

2.2.1 Emergence of the Transformer Architecture

Transformers (Vaswani et al., 2017) represented a breakthrough in sequence-to-sequence modeling tasks, such as machine translation, by relying on self-attention mechanisms which removed the need for recurrent operations. As shown in Figure 2.1, the Transformer is composed of an encoder, which stacks multiple layers of self-attention and position-wise feed-forward networks to encode contextual information for each token, and a decoder, which similarly employs self-attention but also includes a cross-attention layer to attend to the encoder's output. Notably, the self-attention in the decoder is masked to ensure that each position can only attend to preceding positions in the sequence, making auto-regressive generation feasible. This architectural design allows the model to generate sequences step by step, making it suitable for text generation and machine translation tasks. Moreover, the ability to process entire sequences in parallel addresses the bottleneck of recurrent architectures and facilitates large-scale training, a factor that underlies the remarkable success of Transformer-based models beyond machine translation in the field of NLP.

Self-attention is the core innovation of the Transformer architecture. Given an input sequence

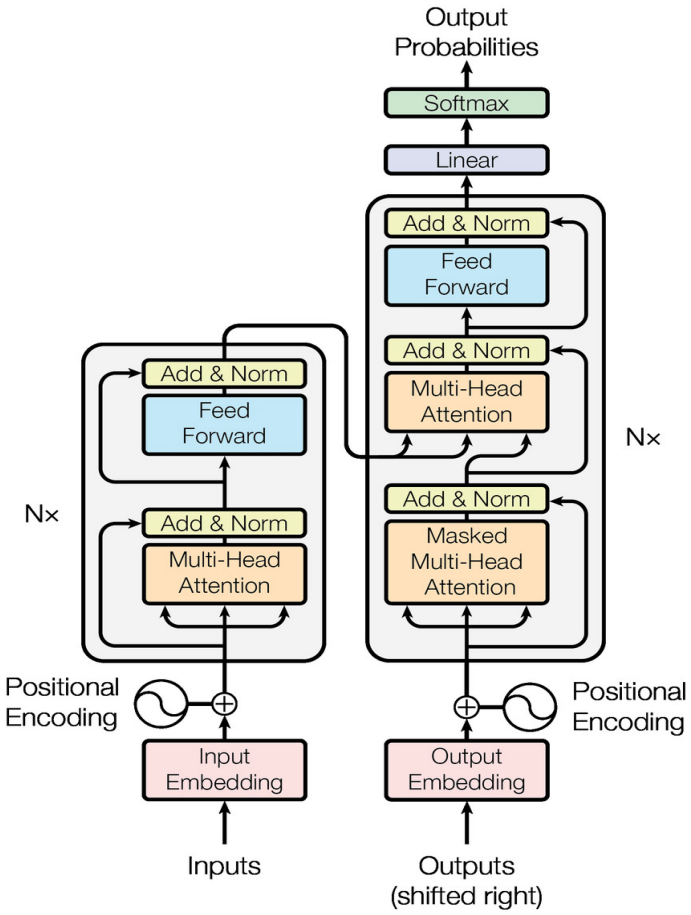

Figure 2.1: The architecture of the Transformer proposed in (Vaswani et al., 2017)

$\mathbf{X} \in \mathbb{R}^{n \times d}$, the self-attention mechanism begins by projecting each token into three distinct learnable spaces—query, key, and value. Formally, these projections can be written as:

$$\mathbf{Q} = \mathbf{X}\,\mathbf{W}^{Q}, \tag{1}$$

$$\mathbf{K} = \mathbf{X}\,\mathbf{W}^{K}, \tag{2}$$

$$\mathbf{V} = \mathbf{X}\,\mathbf{W}^{V}, \tag{3}$$

where $\mathbf{W}^{Q}$, $\mathbf{W}^{K}$, and $\mathbf{W}^{V}$ are trainable weight matrices, and $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the resulting query, key, and value matrices, respectively.

The intuition behind these projections is to determine how strongly each token in the sequence should attend to every other token. Specifically, the alignment scores are computed by taking the dot product between the query and key vectors, scaled by the factor $\sqrt{d_k}$ (the dimension of the query and key vectors). These scores are then normalized via the softmax function and used to weight the value vectors. Mathematically, this is expressed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}\Big)\,\mathbf{V}. \tag{4}$$

Because each element in the sequence can directly attend to all others, this mechanism captures global dependencies in a single pass. Consequently, Transformers achieve highly parallelizable

computations and scalability, outperforming earlier sequence modeling methods. In the next section, we delve into Transformer variants that further refine and extend this foundational architecture for a wide range of applications.

2.2.2 Transformer-based Language Models

Following the success of transformers in machine translation, several studies have explored their application to language modeling, giving rise to three primary branches of models, each leveraging different Transformer components to address distinct linguistic challenges. For example, BERT (Devlin et al., 2018) leverages the Transformer encoder to generate contextualized representations of a sequence, enabling bidirectional understanding of context. In contrast, GPT (Radford & Narasimhan, 2018) uses the Transformer decoder in an autoregressive way, predicting the next token based on preceding tokens, making it particularly effective for text generation and completion. Additionally, models such as T5 (Raffel et al., 2020) utilize the encoder-decoder Transformer architecture, framing all NLP tasks as text-to-text problems, thereby enabling applications such as translation and summarization.

Transformer-based models have consistently demonstrated superior performance across various NLP tasks, including short-text classification, while also reducing the extensive pre-processing required by traditional approaches. Given the focus of the ISCN framework on binary text classification, we opted for models designed for contextual understanding (e.g., BERT) over those optimized for sequence-to-sequence modeling or text generation (e.g., GPT). Figure 2.2 illustrates a straightforward adaptation of BERT for a single sentence classification task by adding a simple feed-forward layer on top of the pre-trained Transformer's output representation.

The models employed in the comprehensive performance analysis study are detailed in Section 4.3.3.

2.3 Bias and Fairness in NLP Systems

ML systems are powerful tools for solving complex problems, yet they can also perpetuate unintended biases. In NLP, these biases may manifest in ways that unfairly disadvantage certain groups, thereby affecting the fairness of downstream models. Systematically studying these biases is crucial for understanding their impact on the fairness of the downstream application and developing effective solutions to mitigate or prevent their harm. However, such research must be rooted in a well-defined conceptual framework with clear motivations and normative reasoning to ensure rigorous analysis and meaningful findings of harmful patterns and devise strategies to address them accordingly within a specified system.

2.3.1 Terminology and Taxonomy

Clear and consistent terminology is fundamental for academic and technical discussions of bias and fairness in NLP systems. With the rapid expansion of NLP-driven applications, establishing a shared vocabulary enables researchers and practitioners to communicate precisely and differentiate nuanced concepts based on their definitions rather than on broad outcomes or evaluation methods (Blodgett, Barocas, Daumé III, & Wallach, 2020). Such clarity further supports reproducibility and helps align methodological approaches in this ever-evolving domain.

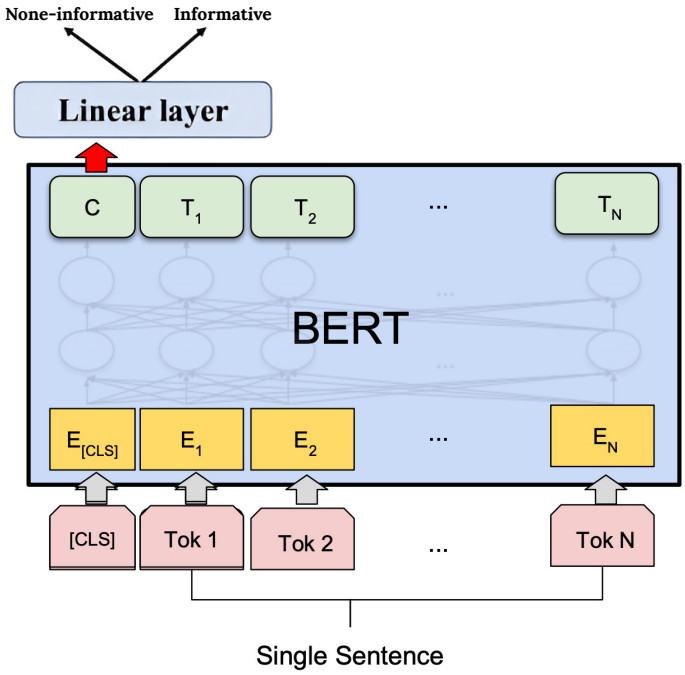

Figure 2.2: Fine-tuning BERT for downstream tasks such as single-sentence classification is as simple as adding a small number of trainable parameters as a classifier and modifying the training objective to minimize cross-entropy loss while leveraging pre-trained representations (Devlin et al., 2018).

In everyday language, "bias" denotes a simple tendency or inclination with negative connotations. In contrast, academic discourse treats bias as a context-dependent concept that varies across disciplines and theoretical frameworks (Hammersley & Gomm, 1997). Within scientific and statistical paradigms, bias is often defined as a systematic deviation from the truth—as seen in the selection or sampling biases that compromise empirical validity of the research outcomes—while a particular focus on social bias in sociocultural analyses extends this definition to encompass deep-seated prejudices and stereotypes rooted in historical and social processes, which in turn perpetuate harmful forms of discrimination against particular social groups (Buolamwini & Gebru, 2018). Recognizing such multiple facets of bias (e.g., technical and sociocultural) underscores the need for each study to tailor its definition of the term to its specific aims and domain.

Within the realm of NLP, and ML in general, bias encompasses multiple meanings that can sometimes appear contradictory. "Inductive bias", for instance, refers to the assumptions that enable a model to generalize beyond its training data, serving as an intentional and necessary component of any learning algorithms (Baxter, 2000). In contrast, "unintended bias" arises when models, including LLMs, inadvertently encode, reinforce, or amplify harmful stereotypes, associations, or disparities that were not part of the model's design objectives (Dixon, Li, Sorensen, Thain, & Vasserman, 2018). Much contemporary work on bias in NLP focuses primarily on this form, particularly with respect to sensitive and protected attributes such as age, gender, and race, and examines how these unintended biases manifest in downstream tasks, including text classification.

In the domain of NLP and LLMs, no single, globally accepted taxonomy for unintended biases exists, though several frameworks are widely used. One common approach distinguishes between

upstream and downstream biases. Upstream biases occur during all processes prior to fine-tuning, including data collection, curation, annotation, and pre-training, reflecting inherent issues in the raw data. In contrast, downstream biases arise during fine-tuning and deployment, potentially introducing new distortions or amplifying existing ones (Steed, Panda, Kobren, & Wick, 2022). Another framework differentiates between intrinsic and extrinsic biases. Intrinsic biases are embedded within a model's internal representations (i.e., word embeddings), while extrinsic biases become apparent in its performance on downstream tasks (Caliskan, Bryson, & Narayanan, 2017). Additionally, explicit bias refers to cases where overt demographic indicators (e.g., "black" or "Muslim") directly influence model behavior, whereas implicit bias arises from subtle linguistic cues, such as style or tone, that correlate with demographic attributes even in the absence of explicit markers (H. Liu, Jin, Karimi, Liu, & Tang, 2021).

Discussing the granular manifestations of bias is a crucial step beyond merely defining it when framing analytical studies. Identity-based bias—a well-known form of explicit bias—has been extensively examined in text classification models and serves as a preliminary lens for bias analysis in emerging NLP applications. In the literature, this bias is also discussed within the framework of spurious correlations, where models learn shortcuts between the training data and task labels (T. Wang, Sridhar, Yang, & Wang, 2021). Its direct impact has been studied in various NLP contexts: in abusive language detection, models may learn discriminatory patterns that harm vulnerable identity groups (Dixon et al., 2018; Nozza, Volpetti, & Fersini, 2019); in gender bias, models may exhibit unequal or stereotypical behavior toward texts containing gender-specific terms (Bartl, Nissim, & Gatt, 2020); and in ableist language, models may inadvertently perpetuate discrimination against people with disabilities (Venkit, Srinath, & Wilson, 2022).

Although defining bias is crucial for analytical rigor, its meaning is inherently entangled with the concept of fairness. As discussed before, bias refers to systematic distortions or prejudices that a model may internalize during training, whereas fairness addresses whether these distortions lead to inequitable outcomes for different users or social groups. In essence, bias concerns the presence of skewed patterns in model behavior, while fairness examines whether these patterns result in inequitable outcomes. It is worth noting that not every instance of unintended bias results in unfair consequences by the model, even though it may still be problematic in its own right—underscoring the need for separate frameworks to assess bias and fairness in ML applications (Dixon et al., 2018). Fairness in ML is often framed through "individual fairness," where similar individuals should receive similar model predictions (Kusner, Loftus, Russell, & Silva, 2017), or "group fairness," where a model's decisions should be statistically independent of sensitive attributes, such as race or gender, that define group membership (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012). Also, a related notion, "subgroup fairness," merges these ideas by enforcing group fairness constraints such as equalizing false positive rates across all subgroups (Kearns, Neel, Roth, & Wu, 2018).

In summary, diverse frameworks have been proposed to characterize bias in NLP and LLMs. These frameworks enable scholars to adopt working definitions that align with their study contexts while distinguishing bias from fairness so that each can be examined independently. In addition, clear articulation of what constitutes bias is crucial because it establishes a shared foundation for exploring its implications, selecting appropriate evaluation methodologies for both bias and fairness, and developing strategies to mitigate or prevent potentially harmful outcomes in NLP systems.

2.3.2 Implications of Bias

The discussion of the implications of unintended bias often involves the inadvertent reinforcement of systemic inequalities, the distortion of decision-making processes, compromised fairness,

and the erosion of user trust in institutions or technologies due to embedded biases that disproportionately disadvantage certain groups—ultimately undermining societal acceptance and legitimacy. These concerns become particularly important in high-stakes applications, where biased outcomes can significantly impact individuals' lives. For instance, in sensitive domains such as automated hiring, algorithmic bias may reinforce discrimination, leading to unfair hiring decisions and workplace inequality (Mujtaba & Mahapatra, 2019; Raghavan, Barocas, Kleinberg, & Levy, 2020); in credit scoring assessments, it can unfairly determine applicants' eligibility for affordable loans (Hurley & Adebayo, 2016); in healthcare, biased clinical decision-support systems and patient risk-prediction models may contribute to disparities in medical treatment (Chen et al., 2023; Obermeyer, Powers, Vogeli, & Mullainathan, 2019); and in legal applications, including automated risk assessments that impact sentencing decisions (Angwin, Larson, Mattu, & Kirchner, 2022; Chouldechova, 2017) and facial recognition systems used in law enforcement (Buolamwini & Gebru, 2018; Garvie, Bedoya, & Frankle, 2016), unintended bias poses risks of unfairness toward different social groups and reinforces existing prejudices, further eroding public trust and fairness.

Given the growing reliance on AI in high-stakes applications, addressing the unintended biases that the backbone models of these systems acquire during training becomes even more urgent, as these automated systems may amplify, rather than mitigate, existing societal biases. Consequently, the potential harm caused by AI-driven systems should serve as the central motivation for rigorous examination of their underlying biases, which necessitates a clear definition of harmful behavior. Establishing this foundation is important because it enables scholars to identify which groups are disproportionately affected by the system, and to understand both the scale and nature of these impacts (Blodgett et al., 2020).

A well-established taxonomy categorizes the harms caused by ML systems into two main forms, namely, allocational and representational harms (Barocas, Crawford, Shapiro, & Wallach, 2017). Allocational harms occur when automated systems allocate resources or opportunities in ways that reinforce existing inequities between social groups. For example, when LLMs are integrated into hiring or university admission tools, they may systematically filter out qualified candidates from historically underrepresented groups due to inherent biases in the training data—biases that reflect and reinforce existing societal prejudices (An, Acquaye, Wang, Li, & Rudinger, 2024; Echterhoff, Liu, Alessa, McAuley, & He, 2024). On the other hand, representational harms occur when systems portray certain social groups in a less favorable light than others, demean them, or neglect to acknowledge their existence. For instance, LLM-generated outputs can omit significant cultural narratives or even reinforce harmful stereotypes; one illustrative example is translation tools that consistently associate particular professions with one gender, perpetuating gender stereotypes (Prates, Avelar, & Lamb, 2020).

It is worth noting that representational and allocational harms are conceptually distinct and can be examined independently. Because allocational harms are often easier to measure and are perceived as having a more immediate impact, researchers—particularly in the domain of LLMs—frequently use them to rationalize the normative importance of addressing representational biases, even when they do not directly measure allocational consequences in their studies (Blodgett et al., 2020). However, this emphasis risks overlooking the subtle yet profound effects of representational harms. Therefore, it is crucial to acknowledge that representational biases, which shape how social groups are portrayed and valued, are inherently problematic, regardless of whether they translate into measurable allocational impacts.

2.3.3 Evaluation and Mitigation Strategies

When selecting an appropriate method to evaluate or mitigate unintended biases in LLMs, it is essential to consider the downstream task context, the type of bias (e.g., intrinsic or extrinsic) to be measured, the structure of the input data available to models, and the nature of the output data generated for measurement.

Several studies indicate that biases in the embedding space exhibit only weak or inconsistent relationships with those observed in downstream tasks. For instance, while Goldfarb-Tarrant, Marchant, Muñoz Sánchez, Pandya, and Lopez (2021) report no reliable correlation, Steed et al. (2022) find that intrinsic and extrinsic biases are somewhat correlated for typical LLMs—though this correlation is largely rooted in the fine-tuning dataset. These findings underscore that bias in representations should not be conflated with bias in downstream applications, thereby recommending a focus on metrics that assess specific downstream tasks (Delobelle, Tokpo, Calders, & Berendt, 2022). Since our work is concerned with applying LLMs to the task of informative sentence classification for customer needs identification systems, we mainly focus on methods assessing and mitigating extrinsic bias.

Identifying and Measuring Bias

In response to the growing awareness of unwanted biases, researchers have developed a range of methodologies and metrics to evaluate unintended bias in text classification tasks, particularly because traditional performance metrics (e.g., accuracy, precision, recall) provide limited insight into how models treat different social groups (Olteanu, Talamadupula, & Varshney, 2017). Evaluating extrinsic bias in downstream text classification often involves analyzing disparities in model performance across demographic groups (Dwork et al., 2012). Following the notion of Equality of Odds—which requires that false positive rates and false negative rates be equal across these groups (Hardt, Price, & Srebro, 2016)—researchers have proposed metrics such as accuracy gaps (the difference in model accuracy between two demographic groups) (De-Arteaga et al., 2019) and the Error Rate Equality difference, which quantifies per-term variation by summing the differences between the overall false positive (or negative) rate and the corresponding per-term rates (Dixon et al., 2018). More specialized, threshold-agnostic metrics have also been introduced, including the Average Equality Gap, which calculates the average difference in correctly identifying positives (or negatives) for a specific subgroup compared to the overall population across all decision thresholds (Borkan, Dixon, Sorensen, Thain, & Vasserman, 2019).

Counterfactual Token Fairness (Garg et al., 2019) is another widely used method proposed for text classification tasks. It is a specific form of counterfactual fairness originally designed to measure individual fairness in causal inference (Kusner et al., 2017) and serves as a complement to the group fairness notion of Equality of Odds. This metric evaluates whether a language model's predictions remain consistent when sensitive tokens are replaced with their counterfactual alternatives. Closely related to this approach, Perturbation Sensitivity Analysis (PSA) is a generic, application-independent framework that detects unintended model biases for named entities in a similar manner, while requiring no additional annotations. Furthermore, PSA-based metrics do not strictly align with individual- or group-based fairness metrics; instead, they assess the perturbation sensitivity of model predictions to score and label shifts across unannotated sentences when substituting names of the same entity type (Prabhakaran, Hutchinson, & Mitchell, 2019).

Mitigating Bias

Mitigation approaches fall into three main categories: pre-processing, in-processing, and post-processing (Friedler et al., 2019). Pre-processing strategies mitigate biases in text classification by modifying or augmenting data prior to training and inference of LLMs. A prominent technique is data balancing and augmentation, where examples from underrepresented groups are synthetically increased, reducing label skew and promoting equitable generalization in LLM classifiers (Dixon et al., 2018). Counterfactual Data Augmentation balances datasets by generating alternative examples that differ only in protected attributes (e.g., gendered pronouns), thereby mitigating bias and ensuring more equitable representation across subgroups (Zmigrod, Mielke, Wallach, & Cotterell, 2019). Removing or downweighting sensitive attributes is another widely used method that falls under the category of pre-processing techniques (G. Zhang et al., 2020).

In-processing mitigation strategies aim to reduce bias by adjusting the learning algorithm or optimization process during LLM fine-tuning. A notable example follows adversarial training concept, in which a secondary classifier attempts to predict protected attributes from intermediate representations, thereby encouraging attribute-invariant features (H. Liu et al., 2021). Another effective method is contrastive learning, which learns representations that bring similar examples closer together while separating dissimilar ones. Chi et al. (2022) introduced conditional supervised contrastive objectives, aligning representations of examples that share sensitive attributes within each task label. Regularization techniques also fall into this category, as they constrain changes in the model's parameters or outputs to minimize spurious correlations while preserving semantic information encoded in the pre-trained model (Chew, Lin, Chang, & Huang, 2024; Nozza et al., 2019).

Post-processing strategies tackle biases after classifier training by directly adjusting predicted labels or scores. Threshold adjustment, for example, modifies decision boundaries for protected groups to achieve fairness criteria. Hardt et al. (2016) exemplify this approach by applying different thresholds and randomization to predictions from logistic regressors. D. Wei, Ramamurthy, and Calmon (2020) proposed transforming predicted probability scores through a function designed to meet fairness constraints while minimizing cross-entropy.

2.3.4 Sources and Root Causes

While we previously discussed pre-training corpora (i.e., mainly from Web text) and downstream datasets, as well as system predictions, as the main potential sources of unintended bias and explored corresponding mitigation methods, it is crucial to also consider systematic factors—such as data collection methodology, task definitions, and annotation guidelines—throughout the model development and deployment lifecycle (Blodgett et al., 2020). For example, Sap, Card, Gabriel, Choi, and Smith (2019) observed that informing annotators about dialectal nuances beforehand leads to a notable reduction in mislabeling tweets written in African-American English as offensive.

In the context of online platforms, participation inequality (commonly referred to as the 90-9-1 rule) further contributes to biased data distributions. According to this rule, approximately 1% of users generate the majority of content, 9% contribute occasionally, and the remaining 90% are largely passive (Ochoa & Duval, 2008). This self-selection bias can result in skewed representations, as the perspectives and behaviors of less active users are often underrepresented or entirely absent.

Having a better understanding of such factors enables practitioners to identify and correct bias at the early stages of model development, effectively preventing its propagation in downstream models.

2.4 Literature Review on Customer Needs Analysis Using User-Generated Content

2.4.1 Customer Needs in Marketing

In marketing, customer needs are defined as the desires, wants, and requirements that drive consumer purchasing behaviors and preferences. To better understand these needs, modeling approaches such as the Kano Model is able to categorize them into three main types: basic, performance, and excitement attributes (Kano, Seraku, Takahashi, & Tsuji, 1984). While satisfying basic needs is essential to prevent customer dissatisfaction, exceeding expectations through performance and excitement attributes plays a crucial role in building strong customer loyalty (Matzler & Hinterhuber, 1998). Moreover, this systematic identification and organization of customer needs, often referred to as the "Voice of the Customer," is a key component of Quality Function Deployment, as it translates customer requirements into tangible design and product specifications (Griffin & Hauser, 1993).

A comprehensive understanding of customer needs is pivotal for numerous applications in product development and marketing. In product development, companies that integrate customer requirements and insights into the design process create offerings that align with market expectations (Law, Majava, Nuottila, & Haapasalo, 2014). Identifying emerging needs also helps them develop new products that increase market share and profitability (Timoshenko & Hauser, 2019). In marketing, consumers can be segmented according to their shared needs which facilitate the use of tailored marketing strategies (Denizci Guillet & Kucukusta, 2016). For instance, companies can implement better-targeted advertising and promotional strategies based on a nuanced understanding of customer preferences of each segment, which leads to increased customer engagement, higher conversion rates, and ultimately revenue growth (Abbas, 2024; Wu, 2023). Furthermore, this understanding enables marketers to identify variations in customers' willingness to pay that are closely tied to perceived brand value and guide the development of effective pricing strategies (Hrinchenko, Robul, & Zalubinska, 2018).

Traditional methodologies for identifying customer needs encompass both qualitative and quantitative techniques, including (but not limited to) surveys, focus groups, interviews, and observational studies. Surveys are a well-known example of quantitative methods used to gather direct consumer feedback, allowing for large sample sizes and statistical analysis (Smets, Langerak, & Rijsdijk, 2013). However, surveys often lack the depth required to capture nuanced customer insights and tend to overlook complex emotional factors. In contrast, qualitative methods such as focus groups and interviews facilitate richer discussions regarding customer motivations. These techniques encompass a variety of tasks, ranging from asking open-ended questions to employing projective methods, which help uncover the underlying and subconscious motivations behind customer preferences—insights that standard surveys cannot capture (Barnham, 2015). Observational techniques offer additional insights by capturing customers' behaviors in real time within their natural environments, enabling the anticipation of evolving needs and the inspiration for breakthrough product innovations (Leonard, Rayport, & Others, 1997). However, these methods can be affected by observer bias and may not fully reveal the underlying reasons for the observed behaviors (Hanski et al., 2014).

In summary, while traditional methods have inherent strengths, their reliance on subjective interpretations, limited scalability, time-consuming nature, and potential biases underscore the need to explore innovative, data-driven approaches to accurately and efficiently capture customer needs.

2.4.2 User-generated Content for Customer Needs Elicitation

With the rapid growth of UGC being stored and shared online every day, there exist an increasing demand for scalable and effective intelligent approaches to identify underlying patterns and extract valuable insights from these extensive collections of data. O'Hern and Kahle (2013) defines UGC as "original contributions that are created by users, are expressed in a number of different media (such as physical objects, sound recordings, computer code, and graphic designs), and are widely shared with other users and/or with firms" (O'Hern & Kahle, 2013). UGC can be found in various forms, particularly in the form of textual data, including microblogs and online reviews.

UGC has fundamentally transformed traditional business models by shifting power from firms to consumers, ushering in a new era of co-creation and customer-driven innovation (O'Hern & Kahle, 2013). The advantages of using UGC for customer needs analysis include access to extensive, rich textual data that is often freely available and continuously updated; moreover, unlike traditional methods such as interviews, professionals can revisit this data for further exploration (Kuehl et al., 2016; Timoshenko & Hauser, 2019). Research has also demonstrated that UGC is a valuable alternative source for uncovering customer needs, yielding insights comparable to those obtained through traditional methods (Timoshenko & Hauser, 2019).

Although UGC offers unprecedented opportunities for firms, its vast and unstructured nature makes manual analysis inefficient and challenging (Salminen, Jung, & Jansen, 2021). Furthermore, the available content is often repetitive or generic (e.g., comments such as "I highly recommend this product."), and tends to focus on a narrow range of customer needs, potentially obscuring rarer insights. These challenges underscore the need for intelligent, and scalable methods to efficiently extract both common and nuanced customer needs from UGC.

2.4.3 Intelligent Methods for Identifying Customer Needs from UGC

Recent advancements in intelligent methods for eliciting customer needs from UGC encompass both semi-automated and fully automated approaches. Semi-automated methods often involve the use of NLP techniques combined with qualitative human analysis, allowing for the extraction of nuanced insights from customer feedback found across platforms like social media and review sites. For instance, Timoshenko and Hauser (2019) illustrate the effectiveness of machine learning to identify relevant UGC content and remove redundancies, facilitating the formulation of customer needs. Similarly, Kauffmann et al. (2019) utilized sentiment analysis to extract and categorize buyer opinions from review data, indicating a trend of integrating human judgment with automated data processing. Fully automated methods frequently leverage advanced machine learning algorithms, including Convolutional Neural Networks and classification frameworks, to parse extensive datasets, eliciting valuable customer insights (Yan, Li, & Fan, 2017). This synergy between extraction and analysis enhances the understanding of customer preferences, providing businesses with valuable insights for strategic decisions and product development (Al Nefaie & Muthaly, 2022; Zhan, Tan, Li, & Tse, 2018).

Despite the advantages presented by these methods, their challenges cannot be overlooked. Automated approaches may suffer from biases in sentiment analysis tools, leading to misinterpretations of customer sentiment (Reshmi & Balakrishnan, 2018). Additionally, reliance on algorithms might overlook contextual factors that significantly influence consumer behavior (Iswari & Putra, 2023; Naeem & Ozuem, 2022a). For instance, while algorithms can effectively gauge overall sentiment, they may fail to capture the subtle complexities of customer dissatisfaction expressed in reviews, resulting in a skewed understanding of consumer needs (Islam, Kaium, Zahan, & Rahman, 2024;

Naeem & Ozuem, 2022b). Comprehensive evaluation of these methods is essential to mitigate such shortcomings, ensuring a balanced representation of customer sentiments and expectations. Bias analysis becomes critical as it helps identify and correct distortions within data interpretation processes, thereby enabling companies to refine their insights and enhance service quality (Y. Zhao & Tang, 2021).

In conclusion, while intelligent methods for eliciting customer needs from UGC are rapidly evolving, their effectiveness relies on a delicate balance between automated processes and human oversight. Continuous evaluation and bias analysis enhance model performance and ensure the insights derived from UGC genuinely reflect customer perceptions and preferences. As the digital landscape becomes increasingly complex, integrating diverse analytical tools will be vital in evolving these techniques to meet customer-centric goals accurately. By addressing the shortcomings of current methodologies and emphasizing the importance of comprehensive evaluations, businesses can better navigate the nuances of consumer behavior in a competitive marketplace (Ana & Istudor, 2019; Chatterjee, Ghatak, Nikte, Gupta, & Kumar, 2023; J. Li & Cao, 2022).

2.5 Leveraging Large Language Models to Advance Customer Needs Elicitation Process

Eliciting customer needs is not a single, isolated task that can be framed as one straightforward ML problem (e.g, ISCN task). Instead, it involves multiple sessions and operations that together form a complete system. For example, (Young, 2004) outlined a 28-step checklist for requirements gathering, encompassing activities such as planning, managing, collecting, reviewing, and tracing. Likewise, (Wiegers & Beatty, 2013) identified 21 best practices for elicitation, which include defining the project scope, identifying stakeholders, reusing existing requirements, modeling the application environment, and assessing the feasibility of proposed requirements. While traditional ML methods can manage certain aspects of customer needs elicitation (e.g., the iscn task for reducing redundancy in UGC), their shortcomings are more pronounced when dealing with the complexities of unstructured data, as discussed in Section 3. Consequently, there remains a need for a more adaptive and fully automated framework capable of capturing structured customer needs and ensuring their feasibility.

Recent advances in LLMs present significant potential to orchestrate the entire customer needs elicitation process, especially for tasks previously difficult to tackle with traditional methodologies. Arora, Grundy, and Abdelrazek (2024) emphasized how LLMs can streamline and automate various stages of requirements engineering, including elicitation, analysis, specification, and validation. Hasso, Fischer-Starcke, and Geppert (2024) introduced a GPT-4-driven approach to generate context-specific questions, thereby improving communication among stakeholders and refining the precision and completeness of requirement specifications. As discussed in Section 3.2.4, eliciting implicit and latent customer needs is a complex task. LLMs can help by creating simulated lead user agents that uncover a wider range of overlooked needs and potential use cases. Notably, empirical findings by Ataei et al. (2024) indicate that mimicking empathic lead user interviews through LLM-based frameworks yields more latent needs than conventional human interviews, underscoring the promise of LLMs in advancing early-stage product development, fostering innovation, and substantially reinforcing the requirements engineering pipeline.

While most existing research focuses on directly extracting customer needs from UGC, the importance of designing an end-to-end system that supports the entire needs elicitation process should not be overlooked. In this regard, an agentic approach offers a powerful framework for leveraging

LLMs to address complex tasks and unify the entire elicitation process. A notable example is the work of C. Wang et al. (2024), who proposed a multi-stage, end-to-end methodology that deploys multiple LLM agents to identify structured customer needs from UGC with minimal human supervision and subsequently convert these findings into actionable product-improvement plans via a feasibility analyzer LLM agent. Yet the approach necessitated human oversight during the feasibility analysis phase to ensure both efficiency and validation, results significantly surpassed earlier techniques in the automated needs elicitation process that relied on traditional unsupervised learning and therefore lacked interpretative and reasoning capabilities (Kuehl et al., 2020).

With generative LLMs demonstrating significant potential in refining and accelerating customer needs elicitation, it is crucial to acknowledge that their integration into production environments entails critical considerations that must be carefully addressed. Proprietary models like GPT variants often come with limited control, uncertain reliability, variable uptime, and higher costs. Hosting large open-weight models locally also demands significant computational resources, making a strategic approach to model usage imperative. In a cost-benefit study, Irugalbandara et al. (2024) observed that smaller LLMs can yield performance on par with larger models on a particular task, offering more consistent results and achieving cost reductions in the range of $5\times$–$29\times$ compared to GPT-4. These findings suggest that for domain-specific tasks, reliance on larger, costlier models may be unnecessary. Furthermore, when multiple small-scale LLMs are incorporated, efficient prompt selection strategies (Y. Liu, Zhang, Li, & Miao, 2024) can serve as a multi-objective optimization approach to balance performance and cost by dynamically selecting the most suitable prompts. By integrating schedule optimization with in-context learning, such a method reduces invocation expenses while maintaining high accuracy, ultimately improving the overall efficiency of LLM-driven systems.

Lastly, it is worth mentioning that conventional supervised learning methods still remain valuable for customer needs identification from UGC due to their cost-effectiveness in large-scale analysis, rapid setup, and simplicity, while the integration of LLMs also opens new possibilities for empowering and enhancing compact-size models such as BERT. One key advantage of LLMs in this context is their ability to streamline data annotation processes by reducing the time and cost required, as discussed in Section 3.2.1. This directly addresses one of the primary challenges of data-hungry supervised learning methods, which often require extensive labeled datasets. Beyond annotation, LLMs can also facilitate targeted data generation, improving model performance on underrepresented or complex subgroups within a dataset. By strategically augmenting training data, LLMs contribute to better model generalization while maintaining overall accuracy (Z. He, Ribeiro, & Khani, 2023). Consequently, a hybrid approach integrating traditional supervised learning with LLM-driven enhancements can create a more efficient and adaptive pipeline for needs identification by retaining the reliability of conventional methods while leveraging language models to overcome data limitations and improve task outcomes.

Chapter 3

A Structured Review of Intelligent Methods Processing Methods for Identifying Customer Needs from User-Generated Content: Challenges, Research Gaps, and Future Directions

In this chapter, we aim to address the research question outlined in Section 1.4: ***What challenges and domain complexities characterize the customer needs (CN) task, how do they affect model robustness, and why should they be incorporated into evaluation frameworks?***.

Abstract

We survey 35 papers proposing intelligent methods for identifying CN from UGC. Our analysis shows that most share generic motivations and objectives, while neglecting the inherent complexity of customer needs and their taxonomies in their evaluation frameworks, despite indications in the literature that such considerations are essential. To highlight this gap, we categorize the surveyed works by their motivations and introduce a critical distinction between using ML as a tool to support CN analysis process in marketing versus treating CN identification itself as a task in ML and NLP. Based on this perspective, we propose three directions for future research: (1) clarifying and consistently defining the task-specific construct (e.g., "needs") to improve transparency and reproducibility, particularly when manual annotation is involved, (2) incorporating the complexity and taxonomy of customer needs into evaluation frameworks, and (3) addressing unintended bias and fairness by situating CN identification within broader social and organizational contexts, thereby ensuring that evaluation frameworks account not only for technical performance but also for equitable and responsible use of NLP systems. These recommendations aim to realign research efforts toward a deeper integration of marketing theory with NLP practice, extending beyond performance metrics to include fairness, transparency, and attention to social context.

Keywords: Customer needs identification, User-generated content, Natural language processing, Customer needs taxonomy, Algorithmic fairness

3.1 Introduction

In marketing, CN represent the underlying desires, wants, and requirements that shape consumer preferences and purchasing behaviors (Griffin & Hauser, 1993). Modeling frameworks such as the Kano model offer a structured view of these needs by classifying them into five categories: basic, performance, excitement, indifferent, and reverse attributes (Kano et al., 1984). While satisfying basic needs is essential to prevent dissatisfaction, and performance as well as excitement attributes are crucial for building strong loyalty, indifferent attributes have little to no impact on satisfaction, and reverse attributes may please some customers but alienate others (Matzler & Hinterhuber, 1998). The systematic capture of these needs, often referred to as the "Voice of the Customer," plays a central role in approaches like Quality Function Deployment, where customer expectations are translated into concrete product and service specifications (Griffin, Price, Maloney, Vojak, & Sim, 2009).

Traditional methods for eliciting CN include surveys, interviews, focus groups, and observational studies (Leonard et al., 1997). Despite their utility, these approaches are time-consuming, difficult to scale, and vulnerable to moderator and observer biases, which can limit depth and reliability (Barnham, 2015; Hanski et al., 2014). At the same time, the growth of UGC—such as online reviews, forums, and social media—has produced a continuous, high-volume stream of customer voice that often matches or surpasses the capacity of manual analysis (Timoshenko & Hauser, 2019) for CN identification. However, the scale, redundancy, and complexity of tone and language of UGC make manual coding impractical, motivating ML/NLP pipelines to filter and structure its insights (Salminen et al., 2021).

Intelligent approaches to eliciting CN from UGC can be grouped into semi-automated and fully automated pipelines. In semi-automated settings, NLP classification methods are applied to remove irrelevant content (Kuehl et al., 2016), clustering and topic modeling organize the filtered data into coherent structures (Zhou et al., 2020), and ranking techniques highlight the most informative items to facilitate human synthesis (Almagrabi, Malibari, & McNaught, 2018). In contrast, fully automated systems often draw on aspect-based sentiment analysis, where product attributes are treated as proxies for CN and negative sentiment toward those attributes is interpreted as an indicator of unmet needs, thereby uncovering customers' wants and demands (Han et al., 2023).

3.1.1 Motivation

CN identification and extraction from UGC is a cross-disciplinary activity, drawing major contributions from computer science, requirements engineering (RE), and marketing. This diversity has led to varying terminology, where concepts such as requirements analysis and customer needs analysis represent overlapping but distinct perspectives. Within this context, several recent works have offered valuable contributions. Salminen et al. (2021) compared manual and automated methods, emphasizing the key challenges each faces alongside their strengths and limitations in requirements engineering. Cheligeer et al. (2022) provided a structured review from seven technical perspectives, ranging from data collection to evaluation and tools. More recently, Cai, Yang, Du, Tan, and Lu (2025) highlighted the unique characteristics of UGC data and product-specific contexts that had been overlooked in earlier reviews, while also proposing a comprehensive V-shaped taxonomy of UGC-based requirements engineering research that serves as a reference point for distinguishing different stages of the process.

Acknowledging that conducting a new systematic literature review in a field of study requires careful assessment of whether existing reviews have already addressed the research questions and

whether a new synthesis can provide added value, we summarized the research questions and contributions of recent literature reviews in the domain of CN analysis in Table 3.1, alongside our work, to provide a comparative view of this study in relation to prior contributions.

While relevant studies have recorded and reviewed the use of ML across different stages of customer needs analysis, we argue that a new perspective is still needed—particularly from the computer science viewpoint, where the literature lacks a structured and critical review. This perspective centers on how existing works in customer needs identification have treated it as a task in ML, a framing that introduces broader methodological and conceptual considerations. Accordingly, this study focuses on that question, and our research questions are formulated on the premise that CN identification should be regarded as a distinct task in the era of ML.

3.1.2 Scope of the Survey

This study focuses on the intersection of three research dimensions: (1) the methodological approach, emphasizing intelligent techniques such as NLP and machine learning; (2) customer needs identification and its subcategories as the primary task of study; and (3) the data source, specifically UGC. Table 3.2 outlines the scope of study according to the taxonomy for literature reviews proposed by Cooper (1988), further clarifying the focus of study and placing its contributions within the broader research landscape.

3.1.3 Organization of the Paper

The remainder of this article is organized as follows. Section 3.2 outlines the key challenges in customer needs identification and the complexities involved in this task. Section 3.3 describes the methodology employed in this study, including the use of ML models in the selection process. Section 3.4 presents the results, and discusses the research questions, broader implications, and social context. Finally, Section 3.5 addresses the study's limitations and provides concluding remarks.

3.2 Review of Challenges and Complexities

Research on intelligent methods for CN identification from UGC is shaped by several challenges that complicate both methodological design and practical deployment. Outlining these factors provides the conceptual foundation for our review and clarifies how challenges in CN analysis intersect with established research problems in ML/NLP constituting a contribution in its own right. An overview of these factors is essential before discussing how our review was conducted and what it reveals.

3.2.1 Data: Terminology, Labeling, and Cost

Terminology

Lack of consistent terminology is one of the main challenges in analyzing UGC for customer needs identification Terms such as "needs," "requirements," "preferences," and even "sentiments" are often used interchangeably across studies, which complicates literature searches, creates inconsistencies in annotation guidelines, and undermines reproducibility. For supervised learning methods that depend on large, well-labeled datasets, this ambiguity introduces significant obstacles, since the meaning of a "need" can shift depending on the study or application domain.

Different authors also adopt different levels of granularity when defining customer needs. For example, in a document-level study, Zhou et al. (2020) labeled reviews as containing needs only if they avoided discussing product attributes or describing how the product fulfilled a need, while excluding reviews of competing brands. Such narrow task-specific definitions may filter out valuable signals, limiting the usefulness of the resulting data for building models that aim to generalize across contexts. To support clarity and comparability, several foundational definitions are drawn from the literature:

- Harrel (1978) divided customer needs into three subcategories: needs, wants, and demands in their pioneering study. In summary, needs refer to the essential requirements for survival, such as water, food, and shelter. Wants, on the other hand, refer to desires that are not necessary for survival, such as a want for a luxury car. Demands are human wants that are supported by the ability and willingness to purchase. For simplicity, the authors considered the term customer needs as a definition that encompasses all three types of needs, wants, and demands.
- Drawing on the definition of customer needs proposed by Brown and Eisenhardt (1995) and Griffin et al. (2009), Timoshenko and Hauser (2019, p. 2) defined customer needs as "an abstract context-dependent statement describing the benefits, in the customer's own words, the customer seeks from the product or service."

Even with such definitions, samples containing implicit needs remain difficult for non-expert annotators. For example, sentences like the second in Table 3.3 include implicit or figurative expressions of needs, which are easy to miss. Excluding such sentences risks overlooking latent needs and biases models toward more explicit, well-represented examples. Consequently, peers are encouraged to provide additional definitions that can simplify the labeling process for non-professional annotators and facilitate the annotation of these intricate samples.

Annotation

Another challenge is the lack of transparency in annotation processes, particularly when datasets are manually labeled by researchers or outsourced annotators. Studies on the ISCN task often fail to clearly define their interpretation of customer needs, a crucial step in reducing ambiguity in the annotation process. Without such clarification, dataset reliability and the feasibility of developing effective ML systems are significantly compromised.

Beyond transparency issues, human labeling is inherently subjective, even when detailed guidelines and precise definitions are provided. For example, Timoshenko and Hauser (2019) and Stahlmann, Ettrich, Kurka, and Schoder (2023) adopted similar definitions of customer needs, yet their labeling diverged in some instances: the sentence *The sound system is great* was labeled as informative in one study, while *It is a great product, I have been using these products for years* was not in the other. These discrepancies pose additional challenges, particularly when models are applied to domains different from their original training context. Furthermore, inconsistent labeling introduces noise into the training data, potentially impairing model learning and degrading performance.

To reduce the influence of subjectivity, one effective approach is to provide detailed and accurate guidelines about nuances of the task (e.g., latent needs). For instance, in the offensive language classification task, informing annotators about dialectal variations significantly reduced the mislabeling of African-American English tweets as offensive (Sap et al., 2019). Also, increasing the number of annotators up to a certain threshold can enhance the agreement score among the annotators (Gamzu, Gonen, Kutiel, Levy, & Agichtein, 2021) alleviating the subjectivity problem.

In addition to refining annotation strategies, implementing post-annotation validation is essential for maintaining data integrity, particularly for datasets manually labeled by researchers. Conducting validation procedures and reporting results can help assess annotation quality and identify inconsistencies. For example, Gamzu et al. (2021) conducted a validation by calculating the internal consistency of the annotated dataset, assuming that similar sentences would have similar helpfulness rates.

In line with the validation process, implementing traditional pre-processing methods such as removing duplicate reviews, excluding short phrases that form incomplete thoughts, and eliminating overly long sentences—often as a result of missing punctuation and containing various other information besides customer needs—can significantly improve data quality.

Cost

CN tasks often require expert knowledge for creating high-quality datasets, especially when using supervised learning methods is involved in the framework. This reliance on expertise comes with substantial costs.

However, alternative approaches such as crowdsourcing can provide reliable labels at scale and at a lower cost than professional annotators (Timoshenko & Hauser, 2019). More recently, the emergence of LLM-based tools has opened new possibilities. For instance, Gilardi, Alizadeh, and Kubli (2023) reported that ChatGPT outperformed crowd workers in tasks like relevance, stance, and topic detection while being substantially cheaper, highlighting its potential as a cost-effective annotation method.

While using LLMs as annotators for creating CN datasets is an innovative idea, naïve reliance on LLM-based labeling does not always improve model generalization (particularly in low-data regimes) and may even degrade performance while introducing hidden computational costs. To address these concerns, recent studies recommend combining LLM-assisted annotation with sampling strategies such as active learning, which prioritizes the most informative examples for the task. This approach helps balance cost efficiency with the need for datasets that meaningfully enhance model performance (Bansal & Sharma, 2023).

3.2.2 Selection Bias in UGC

Customers voluntarily contribute on digital platforms, potentially imposing self-selection bias in UGC when these contributions fail to reflect the broader population's opinions. In the context of online reviews, such bias may manifest in skewed ratings and reviews (Luca, 2015), influencing purchasing decisions both positively and negatively (Xie, Yeoh, & Wang, 2024), and disproportionately representing the needs of certain demographic groups over others.

From a ML perspective, such biases are particularly problematic. Most ML methods assume that training data is an accurate reflection of the target population, an assumption that rarely holds in practice (Fan & Davidson, 2007). When training data lacks representativeness, systematic flaws emerge due to the non-uniform inclusion of instances (Heckman, 1979; Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, & Herrera, 2012; Zadrozny, 2004). Self-selection bias is one of the root causes of this problem in UGC, as voluntary contributions may disproportionately capture specific viewpoints while neglecting others.

As an example, (Kuehl et al., 2016) hold an industry workshop to compile keywords for extracting e-mobility customer needs. While practical, this approach imposed coverage limitations

and induced selection bias by overemphasizing workshop-generated terms, thereby excluding alternative expressions found in natural user feedback. Ultimately, insufficient diversity in training data reduces model robustness and leads to poor generalization on underrepresented inputs (J. Wang et al., 2023).

Yet, even with careful data collection and curation, some degree of overrepresentation is inevitable. Frequently mentioned or easy-to-learn topics tend to dominate training data, causing models to inherit both explicit and implicit biases (Kashi, Lahmiri, & Mohamed, 2025). This often skews model behavior toward majority viewpoints while overlooking the needs of minority groups or those expressed only infrequently or figuratively. For instance, if the term *price* frequently appears in the informative class of a dataset, an ML model may form a spurious correlation between *price* and informativeness, misclassifying non-informative sentences containing the same term. Such biases not only hinder generalization but also raise broader societal concerns, underscoring the importance of actively identifying and mitigating them. Therefore, addressing selection bias requires not only better sampling strategies but also continuous evaluation of model behavior to ensure fair representation of diverse customer needs.

3.2.3 The Influence of Time-Variant Environments on Model Performance

Beyond self-selection bias, another important factor that challenges the assumption of representative training data in ML methods is the temporal evolution of user needs and language. As customer preferences shift, different product aspects gain or lose prominence over time, a phenomenon referred to in ML as covariate shift. Formally, covariate shift occurs when the input distribution changes while the underlying relationship between inputs and outputs remains constant (Moreno-Torres et al., 2012).

In time-variant data streams such as e-commerce platforms, the evolution of input patterns, short-term fluctuations in user interests, and the emergence of new customer needs are prevalent. These dynamics pose additional challenges, particularly for traditional ML techniques that are typically not exposed to such changes during training. Incorporating temporal analyses can help address this issue by capturing how needs emerge, shift, or intensify over time, thereby supporting more robust and context-aware model development.

Despite the relevance of temporal dynamics, only limited work has examined their impact on text classification performance. For example, Agarwal and Nenkova (2022) introduced a method to quantify temporal deterioration and to assess when adaptation is necessary. Their findings show that pre-trained models degrade in the presence of concept shift, where the relationship between inputs and labels changes, but remain relatively stable for tasks like sentiment classification, where label semantics are less time-sensitive. In the context of customer needs, M. Zhang, Sun, Li, Wang, and He (2023) proposed a framework that combines topic modeling with sentiment analysis to assess both initial and supplementary reviews (i.e., those posted long after the original review) to assess temporal shifts in customer sentiments and requirements.

3.2.4 Extracting Implicit Customer Needs from Complex UGC

Finding CN from UGC appears to be more challenging than typical NLP classification tasks due to the abstract, context-dependent nature of consumer demands (Timoshenko & Hauser, 2019). Customer reviews can vary greatly in terms of their content, style, and tone, making generalization difficult for ML models. Unlike observable or explicitly mentioned customer needs, identifying

implicitly mentioned from UGC is not straightforward, as these needs are non-obvious and difficult to articulate since users themselves may not be consciously aware of them (Narver, Slater, & MacLachlan, 2004; Otto & Wood, 2001).

For instance, although all sentences in Table 3.3 convey positive sentiment and demonstrate customers' willingness to utilize the products, only the second expresses a customer need, despite sharing common entities such as *love*, *kid*, and *use*. By finding such needs (which are mainly latent customer needs), companies unlock hidden opportunities to gain a competitive advantage in the market by delivering unexpected value and offering relevant support and services to customers (Bao, Wei, & Di Benedetto, 2020; Timoshenko & Hauser, 2019).

To identify implicit customer needs, Zhou, Jiao, and Linsey (2015) proposed a pioneering NLP approach that combines sentiment analysis with case-based reasoning to extract implicit requirements from online product reviews. Similarly, Han et al. (2023) developed a framework to identify latent needs within the context of figurative language through the extraction of implicit opinions of users by designing a quintuple extraction problem and training a generative LLM in a supervised learning fashion using a manually labeled data set to predict aspect, category, opinion, sentiment, and implicit indicator of a context. Yet the approach and its subsequent enhancement is innovative (Han & Moghaddam, 2024), the manual labeling is pruned to subjective interpretations by annotators leading to unreliable labels, and unstable output of generative LLMs can influence the reliability and consistency of predictions which further add to the level of human intervention in the process.

Identifying implicitly mentioned customer needs is further complicated by factors such as domain-specific jargon, noise, and extraneous information, all of which can negatively affect model performance. Even recently developed state-of-the-art (SOTA) models struggle to handle these complex linguistic features (Potamias, Siolas, & Stafylopatis, 2020), particularly in specialized domains (Lee et al., 2020). To mitigate these challenges, data filtering is commonly employed to reduce the complexity of training datasets. For example, citekuehl2016needmining employed a descriptive coding approach (inspired by (Saldaña, 2021)) to assign categories to segments of 200 randomly selected tweets before labeling the entire database. Subsequently, tweets associated with codes exhibiting low confidence in label correlation were excluded. In one instance, tweets containing URLs were removed because only 11.5% of the sampled tweets contained customer needs. Although this approach effectively reduces data complexity, it risks discarding tweets that may contain previously unrecognized customer needs which are extremely valuable for innovative product design.

Taken together, these challenges highlight that CN identification from UGC cannot be approached as a straightforward classification exercise. Instead, the issues of terminology, annotation, selection bias, temporal variation, and implicit needs must be recognized as defining elements of the task itself. Treating them in this way shifts the focus from developing one-off technical fixes to building evaluation frameworks and methods that reflect the true complexity of the problem. By framing the challenges as integral to the task, future work can develop models and assessments that are not only technically stronger but also more transparent, fair, and aligned with the realities of customer needs analysis.

3.3 Methodology

The review was designed as a structured review rather than a comprehensive systematic mapping, reflecting the targeted scope of the investigation. The primary aim was to capture literature that is both thematically relevant and methodologically aligned with the study's objectives, without extending into unrelated works. The research questions are presented in the Table 3.1.

3.3.1 Planing

The literature review was conducted using a snowballing approach, following the methodological guidelines of Wohlin (2014) and Kitchenham and Charters (2007). We began by selecting a set of seed papers based on the authors' subject knowledge, topic alignment, methodological fit, and the influence of the studies within the field. This strategy increased precision and ensured the inclusion of key contributions that might be overlooked in database searches due to incomplete terminology or indexing gaps. In total, five seed papers were chosen for their direct relevance to ML-based customer needs detection from UGC. These papers, along with the rationale for their selection, are summarized in Table 3.4.

3.3.2 Search Strategy

The utilized search process involved both backward snowballing, in which the reference lists of included papers were examined, and forward snowballing, in which works citing the included papers were screened. These two techniques were applied iteratively, with each newly identified study subjected to the same process, until the last iteration was reached. The snowballing workflow is presented in Figure 3.1.

3.3.3 Inclusion and Exclusion

Following each snowballing iteration, duplicate records were removed using the tools and methods described in section 3.3.4. The remaining records were then subjected to a structured, three-stage screening process applied consistently in both forward and backward snowballing. Screening decisions were guided by predefined inclusion and exclusion criteria (See Table 3.5), accompanied by the authors' judgment where necessary. At the eligibility stage, papers were also excluded if they employed substantially similar methods, particularly when originating from the same research group or authors.

A four-component keyword screening strategy was also applied to filter the obtained papers based on their titles in stage one and abstracts in stage two of the screening process.

- The first component focused on stakeholders, including terms such as "requirement," "need," "preference," "demand," "opportunity," "customer need," "customer preference," "customer requirement," "customer demand," "customer opportunity," "consumer need," "consumer preference," "consumer requirement," "consumer demand," "consumer opportunity," "user need," "user preference," "user requirement," "customer requirement," "user demand," and "user opportunity."

- The second component focused on process or actions related to customer need elicitation, including terms such as "elicit," "eliciting," "elicitation," "identify," "identifying," "identification," "classify," "classifying," "extract," "extracting," "extraction," "discover," "discovering," "mining," "deriving," "gather," "gathering," "capture," "capturing," "inferring," "detect," "detecting," "detection," "acquire," and "acquiring" along with other relevant verb forms.

- The third component addressed UGC sources and analytics, with terms such as "user-generated content," "user generated content," "UGC," "voice of customer," "voice of the customer," "voice-of-customer," "online review," "customer review," "online review," "product review,"

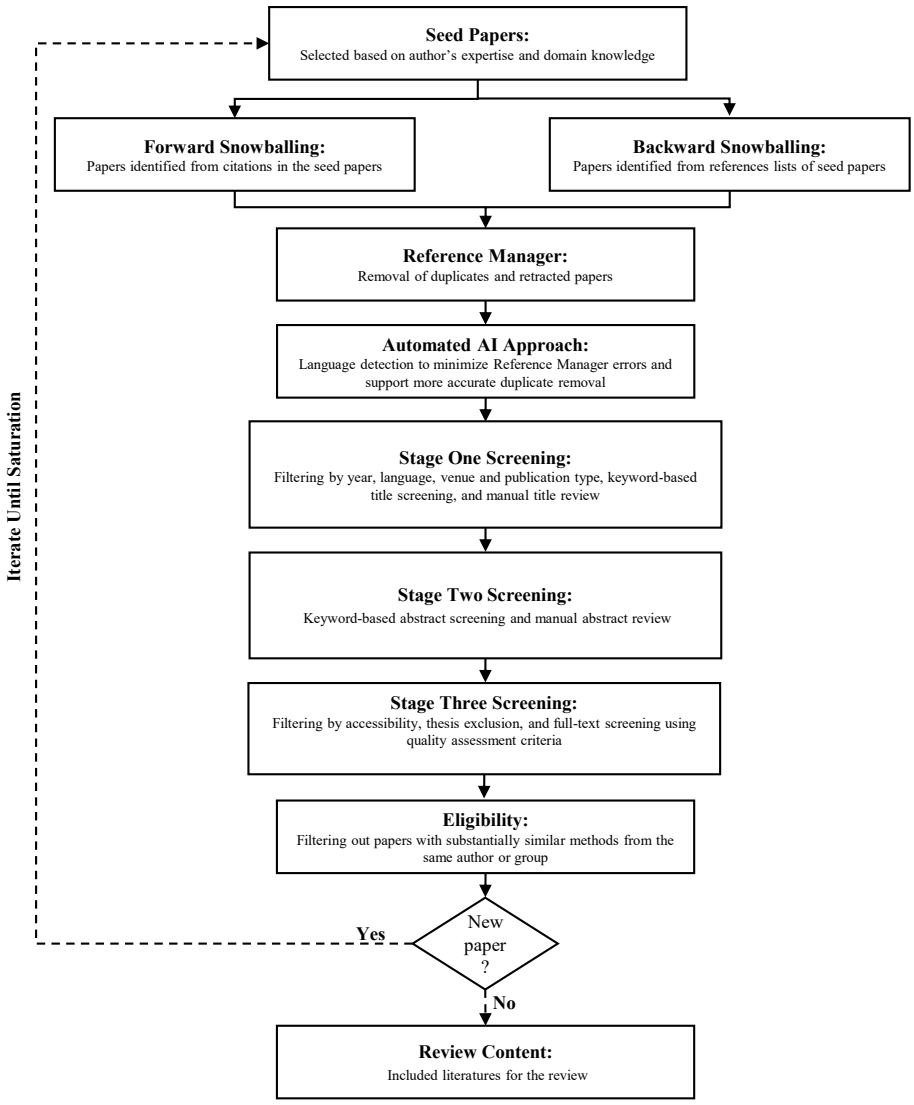

Figure 3.1: A Schematic of the snowballing workflow

"user review," "app review," "application review," "forum," "forums," "social media," "microblog," "microblogs," and "tweet," "tweets," "reddit," "stack overflow," "stackoverflow," "trip advisor," "tripadvisor," "stack-overflow," and "feedback," in addition to descriptors such as "big data," "automated," "data analysis," "text mining," "data mining," "data processing," "data science," "data-driven," "data-oriented," "user-driven," and "decision support."

- The fourth component targeted analytic methods, including terms such as "natural language processing," "nlp," "machine learning," "ml," "deep learning," "transformer," "language model,"

"llm," "bert," "roberta," "gpt," "topic model," "lda," "clustering," "classification," "supervised," "unsupervised," "semi-supervised," "zero-shot," "zero shot," and "artificial intelligence."

In constructing the four-component query, keywords within each component, as well as across components, were combined using the OR operator. In the first stage of screening, a single thematic match was sufficient for an article's inclusion. During the second screening stage, a majority condition was applied: records were retained only if they matched terms from at least two of the four thematic components.

Stage one

The first stage functioned as an initial filter to eliminate clearly irrelevant studies prior to detailed assessment. Eligible works were limited to those published between 2015 and 2025 and written in English. Particular Non–peer-reviewed materials such as keynote talks, tutorials, editorials, and conference abstracts were excluded, although theses and dissertations were retained. Then, keyword filtering was applied using a four-component Boolean query. Subsequently, title screening was conducted by the author to exclude review papers, literature reviews, surveys, and essays.

Stage Two

The second stage applied a more selective filter through abstract-level analysis, combining automated keyword filtering with manual abstract screening by the author. The same four-component keyword query from the stage one was applied.

After automated filtering, remained abstracts were manually reviewed by the author to confirm thematic relevance and exclude borderline cases that did not meet the study's scope. Conversely, studies that showed potential relevance despite incomplete keyword alignment were retained based on the author's judgment.

Stage Three

The third stage combined quality checks with full-text screening. Only papers with retrievable full texts were included, while theses were excluded if corresponding articles by the same authors were already part of the collection.

To assess the quality of the reviewed studies, we applied five criteria: (1) whether customer-needs identification was a primary or clearly defined aim of the study; (2) clarity in task formulation, including definition, granularity, and scope; (3) implementation and evaluation of an empirical ML method rather than purely conceptual discussion; (4) use of UGC as the primary data source with basic source details; and (5) the extent to which contributions, findings, limitations, or implications were clearly articulated. Each criterion was scored as Yes (1), Partial (0.5), or No (0), with equal weights of 0.20, and papers with a total score of at least 0.6 were considered for inclusion.

The final stage was eligibility checking. When multiple papers from the same authors met the criteria, only those with substantially distinct methodological contributions were included.

A detailed summary of the inclusion and exclusion framework, along with examples of excluded papers and the corresponding rationale for each decision, is presented in Table 3.5.

3.3.4 Tools and Software

The processes of reference extraction, citation tracking, and reference management are often tedious and time-consuming, particularly in literature review studies. To address this, a set of specialized tools was employed to support all stages of the snowballing process.

Zotero (*Zotero*, 2007) was used as the primary reference management system to organize retrieved records, store bibliographic metadata, and manage citations throughout the review. Its integration with web browsers and metadata retrieval features facilitated consistent formatting and accurate reference tracking. To enhance data integrity, Zotero was complemented with custom Python scripts for duplicate detection through exact matching. In addition, the Sentence-BERT library (Reimers & Gurevych, 2019) was applied to identify near-duplicate records, and language detection libraries were used to verify the language of paper titles, overcoming Zotero's limitations in this regard.

For forward snowballing searching, Publish or Perish (Harzing, 2016) was employed to retrieve citing papers from multiple databases. For backward snowballing searching, Scholarcy (*Scholarcy*, 2021) was used, leveraging its AI-based reference extraction capabilities to rapidly parse and structure reference lists from full-text PDFs.

To minimize the risk of missing relevant studies due to incomplete metadata, we additionally used Citation Chaser (Haddaway, Grainger, & Gray, 2021), an automation tool that supports collecting both a list of all referenced records, and all citing records based on the Digital Object Identifier of the papers. This significantly accelerated the process and helped capture papers that would otherwise have been missed.

Together, this tool-chain enabled efficient execution of forward and backward snowballing, reducing manual effort and minimizing the likelihood of omitting relevant studies.

3.3.5 Data Extraction

A total of 35 papers that passed the final screening stage were included for data extraction. The quality assessment results for these studies are summarized in Table 3.6.

To improve the transparency and comparability of our study with the corresponding body of work in this domain, we provide a PRISMA-compliant flow diagram (Page et al., 2021) in Figure 3.2.

3.4 Results and Discussions

To start the discussion, we begin with an overview of the corpus as illustrated in Fig. 3.3. The word cloud highlights the most common themes across paper titles, such as "needs," "reviews," "user," "mining," and "language-model," while the yearly distribution of publications shows steady growth in this field of study since 2015. Also, general ideas of all the review papers are summarized and presented in Appendix B.

3.4.1 RQ1: Alignment of Motivations and Contributions with CN Challenges

We categorized the motivations and contributions reported across the 35 reviewed papers into distinct themes. Motivations fell into five categories: (1) limitations of traditional customer CN analysis methods, (2) impracticality of traditional or manual approaches for analyzing UGC, (3) recognition of UGC as a novel source for CN identification, (4) difficulty of detecting implicit, unseen,

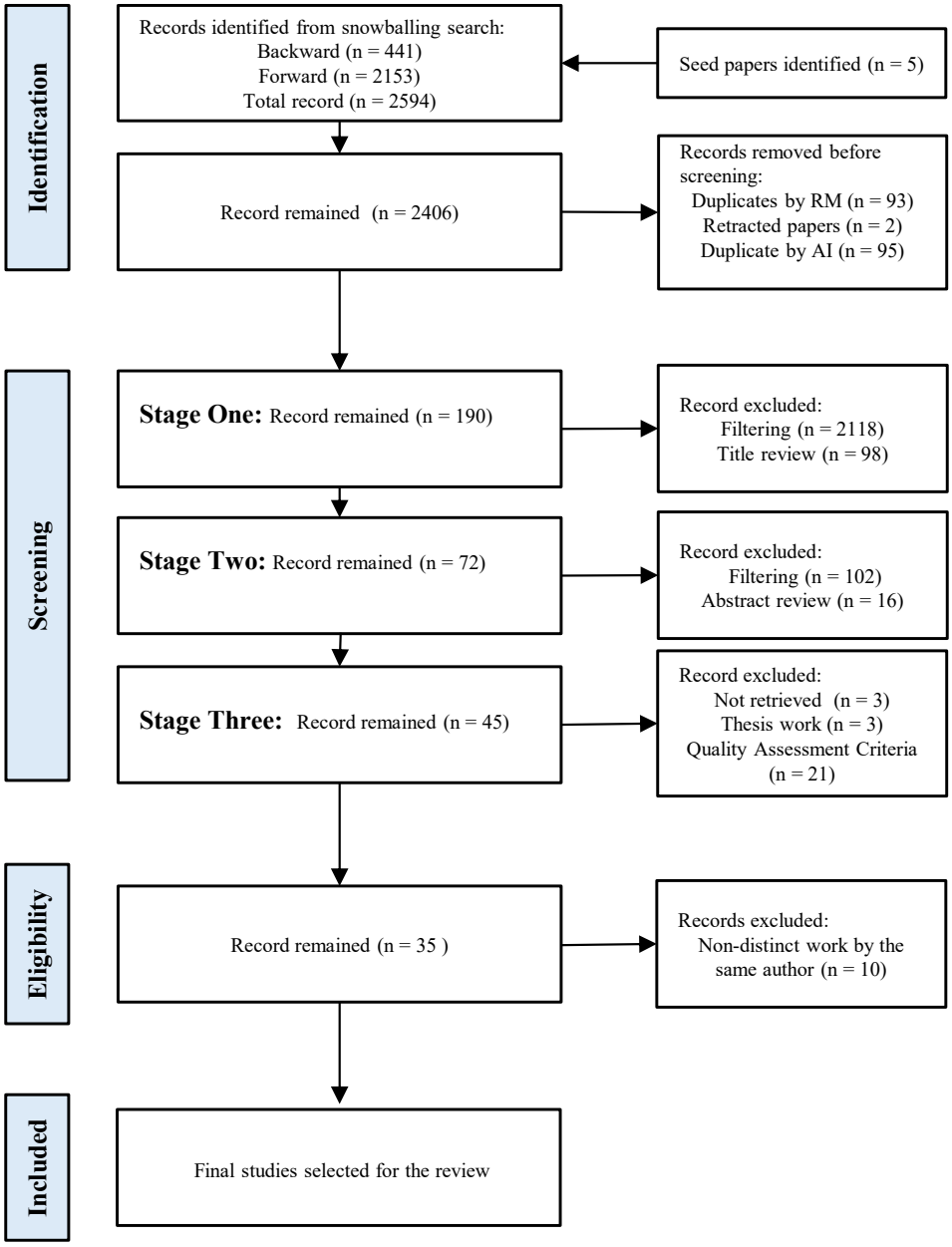

Figure 3.2: A PRISMA flow diagram of article selection

evolving, and future needs in natural language, and (5) opportunities enabled by advanced language technologies such as large language models (LLMs). Reported contributions were grouped into six categories: (1) frameworks or novel CN identification pipelines, (2) comparative and methodological extension studies, (3) introduction of datasets, resources, or benchmark studies (4) novel problem formulations in CN analysis, (5) methods for identifying complex needs (e.g., implicit, infrequent, unseen, or future needs), and (6) LLM-based automation to reduce human involvement.

Across the 35 papers, the most frequent motivations concerned the impracticality of manual

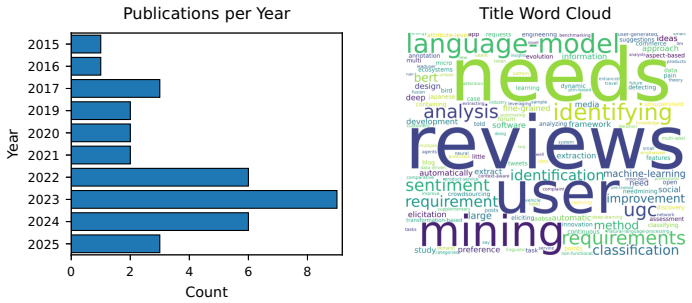

Figure 3.3: Word cloud of paper titles (left) and yearly distribution of included publications (right).

analysis for large-scale UGC (69%), the limitations of traditional approaches (51%), and the challenge of detecting complex customer needs (usually express as implicit, evolving, unseen or future needs) (46%). Other motivations highlighted UGC as a novel source for needs analysis (40%) or opportunities offered by advanced language technologies such as large language models (40%).

The distribution of contributions shows a clear imbalance. Most studies concentrated on developing pipelines or frameworks for customer-needs identification (74%) and on comparative or methodological extensions (68%), reflecting a strong emphasis on incremental technical work. By contrast, fewer papers introduced datasets or benchmarks (14%), proposed novel formulations of the task (25%), or addressed complex needs such as implicit, infrequent, or evolving requirements (20%). Work on automation through large language models was even more limited (11%).

The results point to a clear misalignment. Although complex and evolving needs are often cited as central motivations, only a minority of studies propose methods that directly address them. This imbalance reflects two distinct perspectives on the role of ML in customer-needs analysis. One perspective frames ML as a supporting tool, motivated by challenges such as the impracticality of manual coding, the scale of UGC, and the potential of LLMs. The other treats customer-needs identification itself as an ML/NLP task, as reflected in contributions such as novel task formulations or methods targeting complex needs. Together, these patterns suggest that while the challenges of customer-needs identification are widely acknowledged, they have not yet been consistently translated into methodological works.

3.4.2 RQ2: Incorporation of Complexity and Challenges into Evaluation

Evaluation frameworks in the surveyed studies showed limited engagement with the specific complexities of CN identification. Although nearly all papers relied on manual labeling, only about 35% explicitly defined what constitutes a "need" or referred to an established source, leaving most datasets vulnerable to ambiguity and inconsistency. Temporal dynamics were almost entirely overlooked: just 14% incorporated any form of temporal analysis, and none explicitly operationalized generalization (with a clear definition) as part of their evaluation. Despite frequent claims about the difficulty of implicit, evolving, or infrequent needs, these aspects were rarely reflected in the evaluation design.

Overall, evaluation practices tended to emphasize standard performance metrics on static datasets rather than frameworks that capture the methodological challenges identified in the literature. This narrow focus reinforces the view of ML as a generic tool for text classification, rather than treating CN identification as a distinct ML/NLP task that demands tailored evaluation protocols. Stronger integration of construct clarity, temporal robustness, and generalization checks is required to bring

evaluation in line with the complexities acknowledged by the field. In this respect, positioning CN identification as a task in ML—rather than merely an application of it—remains an opportunity yet to be fully discovered and advanced.

3.4.3 RQ3: Fairness and Social Context in Customer-Needs Identification

None of the surveyed studies explicitly investigated bias or fairness in customer-needs identification. This absence is striking given that overlooking certain groups can have tangible consequences for product design and service delivery. As illustrated by the critique of gender bias in product development (Reuther, 2022), systematic neglect of minority or underrepresented needs can reinforce inequities and limit inclusivity in innovation. In the context of UGC, biases may arise from who contributes feedback, how language varies across demographics, or which needs are most visible in online platforms. Without attention to these dynamics, models risk amplifying dominant voices while marginalizing others.

Integrating fairness and social context therefore requires moving beyond technical accuracy toward evaluation protocols that examine representativeness across groups, track whose needs are being captured or overlooked, and make explicit the social implications of model outputs. Framing customer-needs identification as an ML/NLP task should include fairness diagnostics and mitigation strategies from the outset, ensuring that methods not only perform well but also equitably support diverse customer populations.

3.5 Conclusion

This review synthesized 35 studies on CN identification from UGC and examined how stated motivations translate into methods and evidence. We found a recurring mismatch: papers frequently motivate work by the scale of UGC, the difficulty of implicit and evolving needs, and recent advances in NLP/LLMs, yet evaluations rarely define the "need" construct with precision, seldom test domain or temporal shift, and almost never consider fairness. A central contribution of this review is to reframe CN as an ML/NLP task in its own right—one that demands clear operational definitions, taxonomy-aware evaluation, and socially responsible assessment, not just better pipelines.

This study has limitations. Coverage is representative rather than exhaustive: snowballing search strategy can inherit seed bias. We emphasized qualitative synthesis over meta-analysis, and all three screening stages involved researcher judgment. Some gray literature and non-retrievable texts were excluded.

We see three priorities for future work. First, make task-related constructs explicit: include annotation guidelines, report inter-annotator evidence, and build taxonomy-aware evaluation frameworks that stress both implicit and rare needs. Second, test what matters for practice: standardized protocols for generalization under domain/temporal shift, with explicit out-of-distribution checks. Third, make CN systems equitable and bias-aware by integrating fairness metrics and mitigation across demographics, dialects, and additional relevant subgroups.

In conclusion, by grounding CN in clear constructs, robust and transparent protocols, and attention to social impact, the community can deliver models that surface diverse, evolving, and consequential customer needs—reliably, reproducibly, and equitably.

Table 3.1: Comparative analysis of prior literature reviews and the present study

Paper	Approach	Analysis and Evaluation	Time Span	Research Questions
Salminen et al. (2021)	Integrative	Qualitative	N/A	– How do manual and automated methods differ in detecting customer needs? – How do manual and automated methods differ in detecting customer needs? – What key challenges limit current approaches to customer needs discovery?
Cheligeer et al. (2022)	Systematic	Qualitative	2007-2022	– What elicitation activities are supported by ML? – What data sources build ML-based requirement solutions? – What technologies/algorithms build ML-based elicitation? – What tools support ML-based elicitation methodology? – How to construct an ML-based elicitation method?
Cai et al. (2025)	Semi-Systematic	Qualitative/ Quantitative	N/A	– What UGC data categories are used for automated RE? – What are the methodologies for automated RE? – What requirement representations are derived from UGC?
This Study	Integrative	Qualitative	2015-2025	– How are motivations and contributions of surveyed studies aligned with the identified challenges requirements of customer needs identification? – To what extent do current methods incorporate the complexity and challenges of needs identification into their evaluation frameworks? – Why and how should fairness and social context be integrated into customer needs identification?

Note: "N/A" denotes that specific information was not available in the in the respective study.

Table 3.2: Summary of the focus and characteristics of the study under the literature review taxonomy (Cooper, 1988).

Characteristic	Categories	This study
Focus	Research findings	√
	Research methods	√
	Practices of applications	×
	Theories	×
Goal	Integration	×
	Identification of the central issue	√
	Criticism	√
Perspective	Neutral representation	√
	Espousal of position	×
Coverage	Exhaustive with selective citation	×
	Exhaustive	×
	Representative	√
	Central or pivotal	×
Organization	Methodological	√
	Conceptual	√
	Historical	×
Audience	Specialized scholars	√
	General scholars	√
	Practitioners or policymakers	×
	General public	×

Table 3.3: Examples of sentences with and without customer needs.

Sentence	Label
This is the only product my mother-in-law loves to use for her dentures.	Non-informative
My kids love to use it also!!	Informative
He loves it and feel it really works.	Non-informative

Table 3.4: Summary of Key References on Customer Needs Elicitation

No.	Reference	Title	Strategy	Rationale
1	Griffin and Hauser (1993)	The Voice of the Customer	Forward	Established the terminology and conceptual foundations for customer needs analysis, including their identification and extraction through traditional methods.
2	Timoshenko and Hauser (2019)	Identifying customer needs from user-generated content	Backward/ Forward	Pioneering study in the field of customer needs analysis, presenting a comprehensive end-to-end framework with UGC.
3	Kühl, Scheuren-brand, and Satzger (2020)	Needmining Identifying micro blog data containing customer needs	Backward/ Forward	Widely recognized within the field and frequently cited in the literature.
4	Cheligeer et al. (2022)	Machine learning in requirements elicitation: a literature review	Backward	Comprehensive survey summarizing methods and tasks, primarily from a requirements engineering perspective.
5	Cai et al. (2025)	Automatic requirements elicitation from user-generated content A review of data, methods, and representations	Backward	Recent survey in this field also discussing LLMs and advancements in NLP within the context of customer needs analysis.

Table 3.5: Exclusion and Inclusion Criteria. Any record not meeting the exclusion conditions at a given stage was retained for further screening or final inclusion.

Screening stage	Exclusion reasons
Stage One	– Published before 2015 or after 2025
	– Written in a language other than English
	– Presented in a non–peer-reviewed venue (e.g., notes, talks, etc.)
	– Classified as a review paper, literature review, survey, or essay
	– No match in any of the three keyword components in the title (customer needs elicitation, UGC sources, analytic methods)
	– Title screening by the author
Stage Two	– Abstract does not match at least two of the four related keyword components (customer needs elicitation, UGC sources, analytic methods, process stakeholder data source method)
	– Abstract screening by the author
Stage Three	– Full text not retrievable
	– Full text is thesis work
	– Obtained 0.6 or less based on QAC after full text browsing
Stage Eligibility	– Filtering out papers with substantially similar methods from the same author or group

Table 3.6: Quality Assessment Criteria (QAC)

Reference	Q1	Q2	Q3	Q4	Q5
Kuehl et al. (2016), Timoshenko and Hauser (2019), De Araújo and Marcacini (2021), Han and Moghaddam (2021), M. Zhang, Fan, Zhang, Wang, and Fan (2021), Salminen, Mustak, Corporan, Jung, and Jansen (2022), K. Zhang et al. (2023), Barandoni, Chiarello, Cascone, Marrale, and Puccio (2024), Ettrich, Stahlmann, Leopold, and Barrot (2024), Kilroy, Healy, and Caton (2024), Timoshenko, Mao, and Hauser (2025),	✓	✓	✓	✓	✓
Jhamtani et al. (2015), Ayoub, Zhou, Xu, and Yang (2019), Kovacs, Buryakov, and Kryssanov (2021), Han et al. (2023), C. Wang et al. (2024)	✓	✓	✓	✓	∼
M. Li, Shi, Yang, and Wang (2020), Mahdi, Gupta, Choudhury, and Bansal (2022), Yin, Jiang, Jain, Liu, and Chen (2023), M. Zhang et al. (2023), Kaur and Kaur (2023), Han and Moghaddam (2024)	✓	∼	✓	✓	✓
Kocon et al. (2021), Q. Zhao, Zhao, Guo, Zhang, and Yu (2022), Bian, Ye, Zhang, and Yan (2022), Cong et al. (2023), Q. Li, Yang, Li, and Zhao (2023), Lee, Jeong, Yoon, and Song (2023), Z. Zhang, Dou, Xu, and Tan (2024), Huang, Qin, Chan, and Wang (2025)	✓	∼	✓	✓	∼
Guzman, Ibrahim, and Glinz (2017), Xiao, Li, Thürer, Liu, and Qu (2022), Stahlmann et al. (2023)	✓	✓	✓	✓	✗
W. Wei, Hao, and Wang (2025)	✓	✗	✓	✓	✓
C. Li, Huang, Ge, Luo, and Ng (2018),	∼	∼	∼	✓	∼

✓ = Meets, ✗ = Does not meet, ∼ = Partial.

Chapter 4

Comprehensive Analysis of Transformer Networks in Identifying Informative Sentences Containing Customer Needs

In this chapter, we aim to address the research question outlined in Section 1.4: ***What comprehensive evaluation framework can be developed to determine whether an ML model for customer needs analysis is both sufficiently effective and superior to alternative approaches?***.

Please note that the content of this chapter is based on the work published as follows: Kashi, M., Lahmiri, S., & Ait Mohamed, O. (2025). *Comprehensive analysis of Transformer networks in identifying informative sentences containing customer needs*. *Expert Systems with Applications, 273*, 126785. https://doi.org/10.1016/j.eswa.2025.126785

Author(s): Mehrshad Kashi[1], Salim Lahmiri[2], Otmane Ait Mohamed[1]

[1]Department of Electrical and Computer Engineering, Concordia University
[2]Department of Supply Chain and Business Technology Management, Concordia University

Abstract

The unprecedented rise in user-generated content (UGC) provides businesses with new opportunities to extract customer insights from unstructured data, particularly for identifying customer needs. Intelligent methods offer time- and cost-efficient solutions to extract such insights from the plethora of repetitive and often redundant UGC. However, widespread adoption of these methods faces significant barriers, including high deployment and maintenance costs, data availability challenges, task complexity, and concerns about model efficacy and ethical implications. To facilitate broader adoption of intelligent systems in customer needs analysis, this study evaluates Transformer-based models in terms of generalizability, fairness, robustness, and sample efficiency across various experimental settings to uncover their true performance and identify the root causes of their errors. Our results show that although Transformer-based models improved the F1-score by up to 18% compared to baselines, their limitations become evident when evaluating their performance against task objectives. Key findings include: (i) Transformer-based networks share error

patterns and struggle to identify infrequent or unseen informative samples, (ii) they heavily rely on abundant information and lexical cues for accurate predictions, compromising inter- and intra-domain generalizability, (iii) larger models do not necessarily improve sample efficiency within their domain, and (iv) while optimal cross-domain results arise from complex domain training, adding more in-domain samples does not enhance cross-domain performance. Overall, this research provides crucial insights to help businesses overcome adoption barriers when implementing Natural Language Processing advancements, such as Transformer-based models, in the customer needs analysis process. Source codes are available at https://github.com/mehrshad-kashi/ISCN-UsingTransformerNetworks.

Keywords: User-generated Content, Customer Needs Analysis, Natural Language Processing, Marketing, Transformer Networks, Lexical Bias

4.1 Introduction

With the rise of social media and the vast amount of user-generated content (UGC) available daily, Natural Language Processing (NLP) techniques have recently garnered significant attention as powerful tools to explore and exploit this accessible and ever-expanding source of data. Marketing, inherently a complex field that often relies on human judgment for nuanced analysis (Proserpio et al., 2020), can significantly benefit from Machine Learning (ML) solutions that, at the very least, facilitate large-scale, systematic, cost-effective, and time-efficient exploration of UGC. The insights gained from these explorations aid companies in making better marketing decisions (D. T. S. Kumar, 2020), finding unique product development opportunities (M. Zhang et al., 2021), augmenting and producing innovative ideas for product design and aesthetic processes (Burnap, Hauser, & Timoshenko, 2023), improving products and services (Chang, Yang, & Chen, 2022), understanding customer satisfaction (Aldunate, Maldonado, Vairetti, & Armelini, 2022), and retaining them (de Lima Lemos, Silva, & Tabak, 2022).

In customer needs analysis, utilizing innovative ML solutions is gaining more popularity (Barandoni et al., 2024; Han et al., 2023; Kilroy, Healy, & Caton, 2022; Kuehl et al., 2016; Y. Wang, Mo, & Tseng, 2018). One approach involves identifying Informative Sentences containing Customer Needs (ISCN)[1](Stahlmann et al., 2023; Stahlmann, Ettrich, & Schoder, 2022; Timoshenko & Hauser, 2019), which focuses on distinguishing sentences that convey customer needs from non-informative or redundant content within UGC. Once filtered, marketing experts analyze and exploit user demands from potentially informative content. This elimination process enhances the efficiency of customer needs analysis in terms of both time and cost, given that a substantial portion of UGC is repetitive and redundant.

While the use of AI-based solutions in customer needs analysis is unquestioned due to their scalability and potential to unravel new opportunities, their adoption by companies faces several barriers. These include economic factors, such as the prohibitive costs of ML research, deployment, and maintenance; technical factors, including data availability, task complexity, and model efficacy; and social implications such as discrimination, bias, unintended repercussions, and environmental impact (Cubric, 2020; D. Kumar & Suthar, 2024). Although investigating the societal implications of AI adoption is crucial due to its potential impact on human lives and livelihoods (Hutchinson, Rostamzadeh, Greer, Heller, & Prabhakaran, 2022), this study primarily focuses on the technical

[1]Throughout this study, the term ”informative” specifically refers to content that conveys customer needs. For example, an ”informative sentence” is a sentence that describes one or more customer needs.

barriers as the initial step towards facilitating AI adoption, leaving the analysis of societal impacts for future research.

Despite the use of state-of-the-art NLP models in the ISCN task (Kilroy et al., 2022; Minaee et al., 2021), existing research indicates that their performance lags behind that of well-established binary text classification tasks, such as general sentiment analysis[2] (Csanády, Muzsai, Vedres, Nádasdy, & Lukács, 2024; Raffel et al., 2020). This disparity highlights the complexity of the ISCN task (see Fig. 4.1). We identify four factors, both before and after model training, that contribute to the complexity of this task and its suboptimal performance: (1) limited availability of high-quality, large-scale annotated data due to domain variations, labeling costs, and the need for expert knowledge; (2) the contextual and abstract nature of the task; (3) sample selection bias; and (4) data shifts over time.

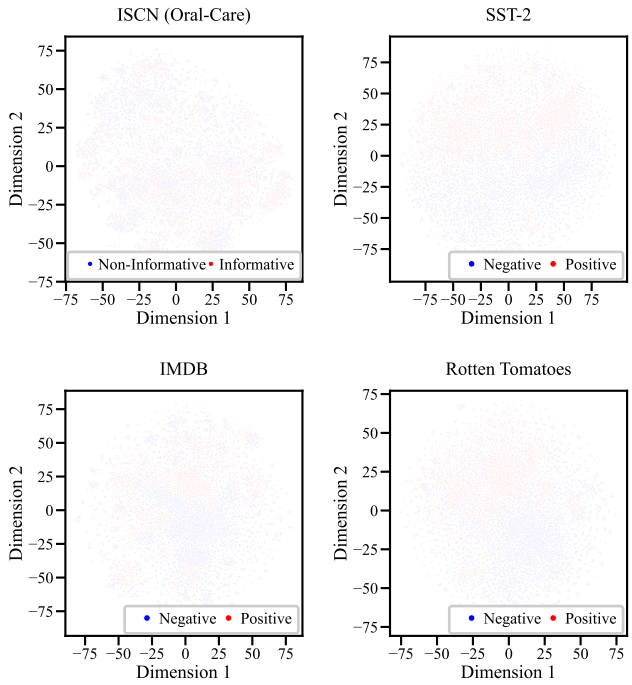

Figure 4.1: T-SNE visualizations of the Oral-Care dataset and 8,000 randomly selected samples from the SST-2 (Socher et al., 2013), IMDB (Maas et al., 2011), and Rotten Tomatoes (Pang & Lee, 2005) datasets. A higher degree of class overlap illustrates the complexity of the ISCN task compared to the selected datasets. Separability Index values are provided in Table 4.6.

There are several challenges associated with this task, including the varying costs of inaccurate classifications due to differing class importance. For example, misclassifying an informative sentence as non-informative (i.e., a false negative) may cause the company to overlook a hidden customer need, whereas mislabeling a non-informative sentence is typically a minor error and often disregarded during manual analysis. Moreover, the semantic distribution of informative sentences is uneven; fundamental customer needs are frequently discussed, while hidden needs are rarely

[2]This does not apply uniformly in all cases, as the complexities of sentiment analysis can vary significantly depending on the context and should not be understated.

mentioned (see the population of needs clusters in Fig. 4.5). Since deep learning models excel at predicting scenarios well-represented in their training data, traditional metrics like accuracy may not sufficiently capture a model's ability to generalize to rarer customer needs or unseen domains. This technical barrier can restrict businesses from effectively leveraging ML solutions, as uncovering and responding to nuanced customer demands is essential. Addressing these demands can aid companies across various stages of innovative product development and design, ultimately securing a competitive advantage for stakeholders in the market(Timoshenko & Hauser, 2019).

To effectively overcome the aforementioned hurdles in AI adoption by companies, it is therefore crucial to deeply understand a model's overall performance. This understanding involves addressing two key questions: 'Is this ML model good enough?' and 'Is this ML model better than alternatives?' To answer these questions, we explore additional evaluation criteria that assess models' ability to handle complex data and accurately identify rare but significant customer needs or samples from new domains. These comprehensive evaluations are important for selecting the most appropriate model, ensuring it excels not just in standard metrics but also in robustness, fairness, and adaptability across various data scenarios.

Building on this foundation, our study evaluates the inter-domain generalizability, robustness to unseen samples, fairness across different need clusters, sample efficiency, and intra-domain generalizability of Transformer-based networks. Through these comprehensive assessments, we aim to identify their strengths and weaknesses from multiple perspectives, particularly in handling challenging dataset subsets with respect to the task objectives, thereby shedding light on their practical effectiveness and limitations. The key contributions of our research are summarized below:

(1) We benchmarked Transformer-based models on the fully-coded Oral-Care domain dataset (Timoshenko & Hauser, 2019), achieving up to an 18% improvement over baseline models.

(2) Through systematic clustering, we analyzed misclassification patterns in the informative class (positive class), grouping samples based on the frequency of their semantic appearances during training. This analysis helped assess the inter-domain generalizability of the models. Additionally, we conducted a dual-setting robustness analysis to gain more granular insights into the models' behavior with unseen informative samples. These findings offer valuable guidance for various stages of model development, such as data selection for customer need analysis.

(3) We evaluated the models' fairness across customer need clusters, uncovering specific weaknesses in Transformer-based models and discussing their implications from a bias analysis perspective.

(4) Using additional datasets curated for the ISCN task, we studied in-domain sample efficiency and intra-domain generalizability. Our findings highlight the advantages of domain-specific training, the effects of dataset imbalance on cross-domain evaluations, and the relationship between domain complexity and predictive accuracy, guiding future model training strategies.

The following section summarizes related research in customer needs analysis and the NLP domain. In Section 4.3, we define the ISCN task, outline the evaluation objectives, and describe our experimental models. Section 4.4 elaborates on the experimental settings, datasets, and evaluation metrics. The subsequent section details our experiments and discusses the results. Finally, We outline the limitations of this study in Section 4.6 before concluding the paper and proposing avenues for future research in Section 4.7.

4.2 Related Work

This section reviews the scientific articles that form the foundation of this study. Section 4.2.1 provides an overview of the current demand for and ongoing efforts to distill helpful information from UGC for eliciting customer needs. Section 4.2.2 presents a brief summary of recent advances in NLP.

4.2.1 UGC analysis using Machine Learning

Unstructured data has significantly contributed to the recent exponential growth of data. According to the International Data Corporation, an estimated 80% of worldwide data will be unstructured by 2025 (King, 2020). This elevates the need and significance of utilizing intelligent methods to effectively process and gain insights from large amounts of unstructured data such as UGC. Utilizing UGC for gaining marketing insights has recently attracted more attention, as discussed in (Berger et al., 2020), which covers commonly used methods, challenges, and potential future directions in this area of research. This work can be considered as part of the ongoing efforts to extract helpful information from UGC, but from a more practical perspective in terms of what needs to be considered to build a practical system using Transformer-based models.

In an early attempt to develop a scalable and intelligent need elicitation process, (Kuehl et al., 2016) analyzed microblog data to identify unmet user needs in the e-mobility domain using three groups of machine learning classifiers to support the design of new products and services. While the SVM model achieved an accuracy of 85%, drew attention to the fact that selecting the best model depends on the project goals pursued by innovation managers, which influences the choice of evaluation metrics. Their findings demonstrated that when the objective is to achieve the highest possible precision (to filter out non-informative content), random forests can achieve high precision (93%). However, this comes at the cost of low recall (4.3%), indicating that only a small number of needs were identified. In a follow-up study, (Kuehl et al., 2020) argued for a scalable ML system to assist companies in the automatic identification and categorization of tweets into predefined customer need groups. However, this approach falls short in identifying samples with rare or unseen needs.

In another pioneering study, (Timoshenko & Hauser, 2019) employed a Convolutional Neural Network (CNN) with domain-adapted embeddings to identify informative sentences from customer reviews, achieving an F1-score of 74%. They demonstrated that professionals could save 45% of their time by sifting only through the informative content filtered by the model. Furthermore, in a comparative analysis, UGC was shown to be a promising alternative data source to traditional methods, such as interviews with potential customers, for identifying customer needs. While the study highlighted the importance of models capable of identifying a broad spectrum of customer needs, it lacked a focus on rare instances, with additional limitations as noted in (Stahlmann et al., 2022).

In a recent study, (Stahlmann et al., 2022) addressed data annotation challenges by using a pre-trained Transformer-based model on a manually labeled multi-domain dataset spanning 32 review categories. They observed a notable improvement in in-domain classification ($\sim$ 6%) compared to traditional deep learning and ML models. While the pre-trained model showed potential performance in cross-domain classification, achieving an F1 score above 70% in 13 of 24 categories, the reliability of these findings is difficult to assess due to the small test dataset sizes in each category and the lack of an analysis to evaluate test sample similarities across all categories. In a subsequent study, (Stahlmann et al., 2023) introduced a multi-domain golden set for benchmarking purposes.

However, our preliminary analysis raises concerns about the dataset's quality, as no rigorous evaluation was performed during the annotation process.

In a comprehensive end-to-end analysis, (Barandoni et al., 2024) investigated the capabilities of both open-source and proprietary large language models in fully automating the identification of customer needs. By employing a few-shot learning approach, the authors addressed common annotation challenges in this domain and demonstrated the potential of large language models when limited or no annotated data is available. However, the transition from a classification task to a text generation task introduced complexities in selecting appropriate evaluation metrics, making it challenging to determine the most suitable model for this purpose. Furthermore, the study did not evaluate the models' abilities to identify latent needs that are not explicitly stated in the text.

4.2.2 Deep learning for text mining

Text mining involves extracting information from unstructured data, and utilizing NLP techniques and ML algorithms to analyze and interpret the content effectively. A critical step in this process is converting text data into numerical vectors. The primary aim of this transformation, which significantly influences classifier performance (Bengio, Courville, & Vincent, 2013), is to retain the semantic and syntactic relationships between words after the transformation. Over time, various vectorization methods have been proposed, including frequency-based representations and neural network models (Bojanowski et al., 2017; Mikolov et al., 2013; Pennington et al., 2014), which primarily excel at word-level representations, along with deep learning-based language representation models (Devlin et al., 2018; Peters et al., 2018).

Neural language models, building upon foundational work in neural probabilistic language models by (Bengio et al., 2003) and later advanced by Transformer architecture (Vaswani et al., 2017), have become a cornerstone of NLP. The Transformer model, originally designed as an encoder-decoder framework for sequence transduction tasks, leverages self-attention to weigh relationships between all tokens in a sequence. By dynamically adjusting word embeddings based on the surrounding context, these models excel in representing polysemous words, syntactic dependencies, and semantic relationships within and across sentences.

A major breakthrough in NLP was achieved with the introduction of the BERT language model (Devlin et al., 2018), which utilized the encoder component of the Transformer architecture to bidirectionally capture information from text. BERT set new benchmarks, surpassing traditional statistical language models and earlier neural network-based approaches by a substantial margin. Notable for its ability to capture contextual relationships, BERT and its derivatives can be pre-trained on self-supervised learning objectives—a resource-intensive process—and subsequently fine-tuned for specific tasks. Fine-tuning (Sun, Qiu, Xu, & Huang, 2019) adapts a pre-trained language model to a target task by optimizing its parameters on task-specific data while retaining the broad, generalized knowledge acquired during pre-training (Rogers, Kovaleva, & Rumshisky, 2021). This strategy underscores the versatility of such models, enabling efficient adaptation to diverse downstream tasks with significantly lower computational costs than training a model from scratch.

4.3 Methodology

4.3.1 Problem statement

ISCN is the task of detecting sentences containing customer needs that appear in the context as explicitly mentioned aspect terms or figurative language (e.g., implicitly mentioned needs). For

example, *"I received this Tablet as my Christmas gift"* is deemed a non-informative sentence, while *"I like the green because it separates mine from my wife's"* implies a need for distinguishable personal items and is therefore labeled as an informative sentence. One of the key challenges in this task arises from customers using implied language, which often does not clearly articulate their needs. This ambiguity makes it difficult for the classification model to identify and accurately classify these demands, which are commonly referred to as latent customer needs.

In this task, each sentence is designated by $S_i = \{t_1, t_2, \ldots, t_l\}$ and class category $Y_i \in \{0, 1\}$, where l is the number of tokens in S_i and Y_i is whether or not the sentence belongs to the informative class. For each informative sentence (i.e., $Y_i = 1$), an associated set of customer needs $N_i = \{n_1, n_2, \ldots, n_r\}$ is defined, where r is the number of identified customer needs within a sentence S_i and n_r denotes a specific customer need from N which contain all the customer needs categories. The primary objective is to develop a rule to predict class labels Y given sentences S in a supervised learning framework using labeled data. This rule is defined by training a Transformer-based model, denoted as $F : S \rightarrow Y$, called a classifier mapping each sentence to its respective class category. Although the existence of need clusters N[3] for sentences is not essential for the training process, their presence facilitates a more comprehensive evaluation of the models.

4.3.2 Evaluation Methods

While prior research underscores the importance of developing scalable and reliable models, the limitations of deep learning models have been less explored. This study aims to bridge this gap by outlining evaluation objectives that consider the task's challenges for comprehensive evaluations of selected models.

Customer needs and requirements constantly evolve, with new needs frequently emerging over time, especially in dynamic environments such as commercial platforms. One of the major challenges in this task is to identify samples with infrequent or unseen needs. Addressing this challenge requires a generic and robust classification model capable of adapting to evolving needs and generalizing effectively to unseen needs.

Objective 1. Let $\text{Sim} = \{\text{Sim}_0, \text{Sim}_1, \ldots, \text{Sim}_x\}$ be a collection of disjoint subsets of sentences from the test set. Each subset, $\textit{Sim}_x$, is a similarity-based cluster that only consists informative sentences whose similar versions have appeared x times during the training process. Two informative sentences are semantically similar if they share at least one customer need. A model is considered robust concerning this objective if it exhibits satisfactory performance across all evaluation metrics during inference over the various subsets of *Sim*. This objective primarily assesses the inter-domain generalizability of models, examining their capabilities in identifying informative samples with diverse frequency levels from less-seen to highly seen samples. The $\textit{Sim}_x$ subgroup is formulated as:

$$\textit{Sim}_x = \left\{ i \in S_{\text{test}} \;\middle|\; \sum_{j \in S_{\text{train}}} \delta(i, j) \;=\; x \right\}, \tag{5}$$

where S_{train} and S_{test} are training and test sets, respectively, x is the similarity value that denotes the exact number of samples in S_{train} that are similar to each sample in $\textit{Sim}_x$, and $\delta(i, j)$ represents the

[3]Customer needs clusters can be identified through both manual annotation and automated methods, such as clustering algorithms.

similarity function between two sentences of training and test sets and is defined as:

$$\delta(i,j) = \begin{cases} 1, & \text{if } |N_i \cap N_j| \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Given the formula, the performance of models on the subset Sim_0 (samples with similarity value of 0) reveals their robustness against unseen informative samples.

Additionally, some customer need categories are more prevalent and frequently discussed, while others are less common, resulting in an inherent imbalance in the dataset. This distribution naturally occurs and is preserved during data collection and preparation unless explicitly adjusted, as it reflects real-world distributions. Therefore, ensuring consistent and high performance across all customer need clusters, regardless of the frequency of a specific cluster, is an essential objective.

Objective 2. Let $C = \{C_{n_1}, C_{n_2}, \ldots, C_{n_m}\}$ be a collection of subsets from the test set, where each subset C_{n_m} consists of informative samples mentioning the customer need n_m and m is the total number of need clusters in the dataset. C may not necessarily comprise disjoint subsets since informative samples can encapsulate multiple customer needs, allowing them to be members of several customer need clusters. Each cluster can be viewed as a collection of samples representing a specific customer need. A fair model is expected to demonstrate consistent and high performance across all customer need clusters in C during the inference stage, as evaluated by relevant metrics. This objective implies that the model needs to detect all types of customer needs with equal and high efficacy.

Given the difficulties and costs of obtaining labeled data for new domains, developing a sample-efficient system for the ISCN task is of great importance. The model also needs to adapt to new domains with fewer cross-domain labels rather than starting from scratch. This approach is known as supervised transfer learning in literature.

Objective 3. Let D_s represent the source domain, and $D_t = \{D_1, D_2, \ldots, D_T\}$ signify the target domains for the same task, where T denotes the number of target domains. Assuming a network is initially trained on D_s, The objective is to derive a model that can harness the knowledge from D_s and adapt to a target domain D_i (intra-domain generalizability), either label-free or with minimal labels from D_s.

Fig. 4.2 presents a flow diagram that illustrates the primary steps in developing a customer needs identification system, with an emphasis on the main focus of this study: a comprehensive evaluation of Transformer-based models. It is also important to note that the social implications of any AI-based system are crucial to consider before real-world implementation, an analysis that we reserve for future study.

4.3.3 Models

In this study, we employed various Transformer-based models (Vaswani et al., 2017), which vary in size, architecture, and training strategy, and have demonstrated outstanding performance in complex NLP tasks. Unlike traditional neural network models such as LSTM and CNN, which lack an integrated embedding layer, Transformer-based models utilize a network of fully connected tokens enhanced by the self-attention mechanism. This allows a direct exchange of information across all positions, making them a suitable choice for the ISCN task, considering the quality and quantity of the available data for training. Table 4.1 provides a detailed summary of the models utilized in this study.

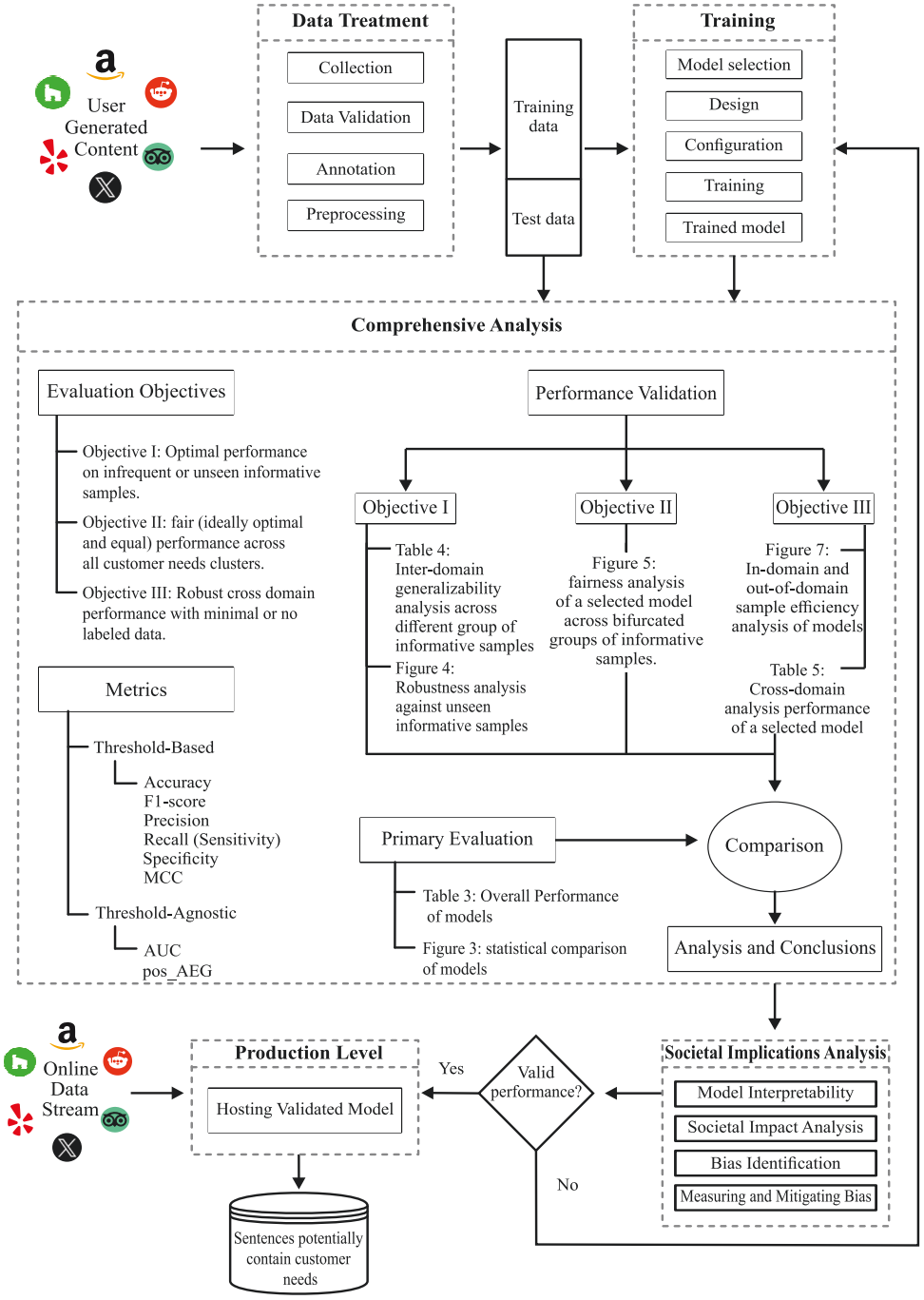

Figure 4.2: Overall flow diagram of the ICSN task. The Objective-based In-depth Evaluation block is the primary focus of this study, while the Societal Implications Analysis and Production Level blocks are beyond its scope.

Table 4.1: An overview of the Transformer-based models utilized in this study, highlighting their core concepts, parameter counts for selected variants, training objectives, pre-training datasets, and macro-average GLUE benchmark scores.

Name	Main Ideas	Size	Training Objectives	Pre-training Corpus	GLUE Score	Reference
BERT	Revolutionized NLP by pre-training deep bidirectional representations using the Transformer encoder, leveraging two novel self-supervised objectives along with massive training data.	Base: 110M Large: 340M	Mask Language Modeling (MLM), Next Sentence Prediction	BooksCorpus, English Wikipedia	84.05	(Devlin et al., 2018)
XLNet	Built on the Transformer-XL decoder, which used segment recurrence and relative positional encoding to handle long-range dependencies, it employed permutation-based language modeling to capture bidirectional context within an autoregressive framework.	Base: 110M Large: 340M	Permutation Language Modeling	BooksCorpus, English Wikipedia, Giga5, ClueWeb, Common Crawl	89.15	(Yang et al., 2019)
RoBERTa	Enhanced BERT through better hyperparameter tuning, removal of NSP from training process, and training on considerably larger datasets with longer sequences.	Base: 125M Large: 355M	Dynamic MLM	BooksCorpus, English Wikipedia, CCNews, Open WebText, STORIES	89.42	(Y. Liu et al., 2019)
Distil-BERT	Compressed BERT into a smaller, faster model via knowledge distillation while retaining 97% of its performance.	66M	MLM with Distillation	BooksCorpus, English Wikipedia	79.60	(Sanh, Debut, Chaumond, & Wolf, 2019)
Distil-RoBERTa	Compressed RoBERTa, using knowledge distillation with the same strategy as DistilBERT, and trained on a reduced corpus compared to the original RoBERTa.	82M	Dynamic MLM with Distillation	Open WebText	82.35	(Sanh et al., 2019)
ALBERT	Optimized BERT architecture by reducing memory usage and accelerating training by significantly decreasing the number of parameters through parameter sharing and factorized embeddings.	Base: 11M Large: 18M	MLM, Sentence Ordering Prediction	BooksCorpus, English Wikipedia	84.75	(Lan et al., 2019)
DeBERTa	Enhanced BERT by separating content and positional information, utilizing disentangled attention mechanisms, and improving the masked token decoder by integrating position embeddings prior to predictions.	Base: 100M Large: 350M	MLM with enhanced mask decoder	BooksCorpus, English Wikipedia, Open WebText, STORIES	89.87	(P. He, Liu, Gao, & Chen, 2020)
XLM-RoBERTa	Multilingual RoBERTa trained on 2.5 TB of data across 100 languages.	Base: 270M Large: 550M	Multilingual Dynamic MLM	Multilingual subset of Common Crawl	86.21	(Conneau et al., 2019)

Table 4.2: Summary statistics of the sentence-level product review datasets used in this study.

Statistic	Oral-Care		Electronics		Baby		Sports-Outdoors		Pet Supplies	
	non-inf.	inf.	non-inf.	inf.	non-inf.	inf.	non-inf.	inf.	non-inf.	inf.
Number of samples	3819	4181	1214	786	1144	856	1262	738	1361	639
Number of tokens	73640	107020	17771	16109	13946	14277	12097	10823	16686	11551
Number of unique tokens	6462	7199	2812	2650	2226	2048	1985	1823	2565	1863
Number of stop-words	4020	4871	1104	767	890	702	922	642	1081	571
Number of unique stop-words	211	204	129	109	119	108	110	90	133	92
Average length	19.28	25.59	14.63	20.49	12.19	16.67	9.58	14.66	12.26	18.07

Despite the versatility of pre-trained language models, domain-specific language can reduce their performance (Ma, Xu, Wang, Nallapati, & Xiang, 2019; Peng, Chersoni, Hsu, & Huang, 2021). Domain adaptation enhances model performance by tailoring pre-trained models to specific task domains, thereby improving their ability to interpret domain-specific language. To demonstrate the effectiveness of domain adaptation, we utilized checkpoints from a study in which $\text{RoBERTa}_{\text{base}}$ was pre-trained on Amazon reviews (Gururangan et al., 2020). This approach not only showcased domain adaptation's efficacy in improving performance but also contributed to reducing our carbon footprint by reusing existing models rather than training new ones from scratch. We refer to this model as $\text{RoBERTa}_{\text{base+DA}}$ throughout this paper.

4.4 Experiments

4.4.1 ISCN Datasets

We conducted primary experiments on the dataset employed in (Timoshenko & Hauser, 2019). This dataset[4], comprising eight thousand fully coded sentences, originates from Amazon Oral-Care product reviews (R. He & McAuley, 2016) published between 1994 and 2014. Three professional analysts from a marketing company carefully annotated the dataset by identifying the specific customer need(s) addressed in each sentence, yielding 82 distinct customer need groups (i.e., N=82). The availability of need clusters and professional annotations makes this dataset an excellent resource for our comparative studies.

We also incorporated four additional datasets introduced by (Stahlmann et al., 2023) for the cross-domain and sample-efficiency experiments. Each dataset contains sentences from product reviews of three top-selling Amazon products in the categories of Electronics, Baby, Sports-Outdoors, and Pet Supplies and is annotated by three annotators.

Table 4.2 presents the numerical characteristics of all datasets, categorized by their domains. The table indicates that the Oral-Care dataset exhibits a nearly balanced distribution of class labels, in contrast to the unbalanced distributions observed in the other categories. Variations in these distributions may stem from label discrepancies and having different levels of sensitivity to the definition of customer needs in the annotation process. It is important to note that the same definition of customer needs was applied across all five datasets.

[4]The dataset is not publicly available. Interested readers may contact its authors to request access.

4.4.2 Evaluation Metrics

To obtain a quantitative understanding of errors across our models during the primary evaluation, as discussed in Section 4.5.1, we employed standard performance metrics for classification, including accuracy, precision, recall, and F1-score. The formulas for these metrics, calculated for the informative class, are presented below:

$$Accuracy = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{7}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{10}$$

$$F_1\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

where TP represents the number of true positive results, TN stands for true negatives, FP refers to false positives, and FN refers to false negatives. We further computed the Area Under the Curve of the receiver operating characteristic, which shows the performance of the classifier across all T values for the informative class ($AUC = \int_0^1 ROC\, dT$ where T is the decision threshold). This metric provides a qualitative evaluation of a model's ability to accurately rank examples of both classes.

To evaluate the impact of lexical bias on model predictions in Section 4.5.2, we used a score-based metric from (Borkan et al., 2019). This metric offers several advantages over traditional metrics such as accuracy. It is scale-invariant, robust against imbalanced groups, and provides detailed insights into model performance across specific subgroups. Essentially, it helps to understand how models process sentences containing specific linguistic cues compared to sentences within the same class that lack those cues, thereby assessing an additional aspect of model performance from a fairness perspective.

Positive Average Equality Gap (pos_{AEG}) evaluates model fairness in predicting informative samples by comparing the separability of prediction score distributions across various informative subgroups. Positive data points are selected randomly from two distributions, the subgroup, and the background, expecting an equal probability for either prediction score to be higher. Mathematically, it is defined as:

$$pos_{\text{AEG}} = \frac{1}{2} - P\{Y_i > Y_j | Y_i \in D^+_{BG}, Y_j \in D^+_{SG}\}. \tag{12}$$

This metric ranges between -0.5 and $+0.5$, with zero representing optimal performance. In other words, this metric seeks to find low separability between subgroup and background samples, both of which belong to the informative class. A negative pos_{AEG} indicates a leftward shift in positive sample predictions, potentially increasing false negatives. The reversed definition applies to the neg_{AEG} metric.

In the experiments on sample efficiency and cross-domain adaptation, as detailed in Sections 4.5.3 and 4.5.3, we employed the Matthews Correlation Coefficient (MCC). This metric is particularly well-suited for binary classification tasks because it assigns high scores only when the model performs well across both informative and non-informative classes (Chicco & Jurman, 2020). Its sensitivity to class balance makes it an ideal choice for out-of-domain evaluations, where the metric must accurately reflect the performance disparities caused by class imbalances.

$$\text{MCC}_i = \frac{\text{TP}_i \times \text{TN}_i - \text{FP}_i \times \text{FN}_i}{\sqrt{(\text{TP}_i + \text{FP}_i)(\text{TP}_i + \text{FN}_i)(\text{TN}_i + \text{FP}_i)(\text{TN}_i + \text{FN}_i)}} \tag{13}$$

4.4.3 Training Details

We implemented and fine-tuned our models using Hugging Face library (Wolf et al., 2020), largely following the hyperparameter settings reported by (Devlin et al., 2018). The *[CLS]* token was utilized for classification, processed through a linear layer with a *Tanh* activation function and a 10% dropout rate. We fine-tuned the models using learning rates of 1×10^{-5} and 2×10^{-5} for large models, and 3×10^{-5} for others. The *AdamW* optimizer was employed, with a weight decay of 0.01 and a linear scheduler set to a warm-up ratio of 10%. All models were trained with a batch size of 16 for up to 10 epochs, usually stopping around the fifth epoch due to an early stopping strategy. In the sample efficiency experiment, we reduced the batch size as the number of training samples decreased, reaching a final size of four. Due to potential mispunctuation, the length of input sentences can vary; therefore, we limited the maximum number of tokens per sentence to 40, accommodating more than 95% of training samples and minimizing excessive padding.

4.4.4 Evaluation

In the benchmarking experiment in Section 4.5.1, we used 10×5 cross-validation approach, meaning ten different repetitions of 5-fold Cross-Validation. Random seeds for repetitions were: $\{94, 791, 5, 6932, 1759, 323, 1694, 9741, 200, 999\}$. Each time, 80% of data (about 6.4 thousand sentences) was selected for training and 20% (around 1.6 thousand sentences) for testing. We avoided preprocessing to preserve language structure and enable fair comparisons across methods, highlighting the capability of Transformer models to handle diverse inputs. Experiments were conducted on a 40GB A100 GPU. Details for other experiments are in their relevant sections.

4.5 Results and Discussion

4.5.1 Informative Sentence Classification Results

Table 4.3 presents the results of Transformer-based models as well as baselines from (Timoshenko & Hauser, 2019) on the Oral-Care domain of the ISCN task. All Transformer-based models consistently outperform baselines.

Generally, the larger versions of the models perform better than their base versions, except for $\text{XLNet}_{\text{large}}$ and $\text{ALBERT}_{\text{large}}$, which do not perform as well as their base versions[5]. RoBERTa

[5]We found fine-tuning of $\text{ALBERT}_{\text{large}}$ and $\text{XLNet}_{\text{large}}$ very challenging with a unified setting for all training repetitions, which may affect the reliability of their results. These results are included to ensure a comprehensive evaluation and should be interpreted cautiously.

Table 4.3: Accuracy, F1-score, Precision, Recall, and AUC of Transformer-based models for the Oral-Care domain of ISCN task. Values are presented as mean ± standard deviation. The highest and second-highest scores across all models are highlighted in bold and denoted by * and **, respectively.

Model	Accuracy	F1-Score	Precision	Recall	AUC
Baselines					
SVM	64.60	65.70	63.70	67.90	
LSTM	73.20	73.40	72.80	74.00	
CNN	74.20	74.00	74.40	73.60	
CNN (asymmetric cost)	70.00	74.00	65.20	85.30	
This Study					
DistilBERT	80.47±0.76	81.75±0.81	79.91±1.19	83.70±1.44	88.64±0.68
DistilRoBERTa	81.73±0.76	82.92±0.75	81.10±1.47	84.87±1.64	89.84±0.61
$\text{ALBERT}_{\text{base}}$	81.15±0.86	82.11±0.94	**81.60±1.43***	82.84±1.59	88.52±1.04
$\text{BERT}_{\text{base}}$	81.44±0.81	82.44±0.80	**81.56±1.35****	83.36±1.37	89.08±0.57
$\text{XLNet}_{\text{base}}$	80.80±0.90	82.26±0.84	79.52±1.43	85.22±1.43	88.87±0.76
$\text{RoBERTa}_{\text{base}}$	82.48±0.88	83.69±0.86	81.50±1.38	86.03±1.39	90.11±0.75
$\text{RoBERTa}_{\text{base+DA}}$	**82.87±0.96****	**84.18±0.83****	81.37±1.54	87.24±1.49	**90.54±0.78****
$\text{DEBERTa}_{\text{base}}$	81.92±0.98	83.28±0.94	80.58±1.49	86.20±1.29	89.48±0.70
$\text{XLM-RoBERTa}_{\text{base}}$	81.49±0.98	82.90±0.96	80.14±1.52	85.88±1.39	89.51±0.83
$\text{ALBERT}_{\text{large}}$	79.54±0.82	80.55±0.89	80.03±1.47	81.11±1.48	86.59±0.83
$\text{BERT}_{\text{large}}$	81.59±0.76	82.75±0.85	81.06±1.34	84.56±1.75	88.94±0.83
$\text{XLNet}_{\text{large}}$	79.53±1.26	81.19±1.19	78.11±1.74	84.57±1.83	87.10±1.05
$\text{RoBERTa}_{\text{large}}$	**82.92±0.90***	**84.25±0.84***	81.30±1.40	**87.46±1.35***	**90.90±0.68***
$\text{DEBERTa}_{\text{large}}$	82.62±0.87	84.00±0.90	80.96±1.53	**87.32±1.56****	89.88±1.00
$\text{XLM-RoBERTa}_{\text{large}}$	82.24±1.00	83.62±0.93	80.71±1.54	86.79±1.69	90.07±0.78

variants outperform their corresponding versions in all other model families. Notably, $\text{RoBERTa}_{\text{large}}$ tops the list among all models with an F1-score of 84.25%, representing a 10% improvement over the best baseline result. In addition, RoBERTa models are also superior choices for identifying informative sentences considering the recall metric.

In terms of parameter efficiency, $\text{ALBERT}_{\text{base}}$ is six times smaller than DistilBERT, yet it surpasses it in F1-score and nearly matches the F1-scores of $\text{XLNet}_{\text{base}}$ and $\text{BERT}_{\text{base}}$ with almost ten times fewer parameters. This performance makes $\text{ALBERT}_{\text{base}}$ an ideal candidate for scenarios with limited computational resources, and given its superior precision among all models, it is a top choice for scenarios demanding high precision for this task.

Regarding domain adaptation efficacy, the $\text{RoBERTa}_{\text{base+DA}}$ model, with at least three times fewer parameters than the best-performing model, $\text{RoBERTa}_{\text{large}}$, and other large models, ranks a close second in our study, achieving an F1-score of 84.18%. This represents a 0.5% improvement over its non-adapted variant, underscoring the value of domain adaptation for the ISCN task, consistent with the findings reported in the literature(Gururangan et al., 2020).

Statistical Analysis

We statistically compared the performance of all models using multiple paired t-tests to ascertain the significance of minor differences. We hypothesized that the mean difference between the two method results is zero in each pairwise comparison. However, metric results are correlated across folds in repeated k-fold cross-validation for two deep learning models. Consequently, a pairwise comparison without adjusting for this correlation underestimates the variance, increasing the risk of Type I errors (false-positives) in the test (Dietterich, 1998). Denoting n_1 and n_2 as the number of training and test samples, respectively, to ensure the validity of the independence assumption of the paired t-test, we computed the corrected t-statistic for multiple tests using the formula proposed by

(Nadeau & Bengio, 1999):

$$t_{\text{stat}} = \frac{\mu_N}{\sqrt{\frac{1}{N} + \frac{n_2}{n_1}\sigma_N^2}} \tag{14}$$

where μ_N, σ_N^2, and N represent the mean, the variance of the pairwise differences, and the number of experiment runs (50 in our case), respectively.

The adjusted p-values are presented in Fig. 4.3. Considering the F_1-score as the primary metric, the differences between the large versions of RoBERTa, DeBERTa, and XLM-RoBERTa are statistically insignificant. In contrast, for the AUC metric, $\text{RoBERTa}_{\text{large}}$ shows significantly better performance compared to all other models, except $\text{RoBERTa}_{\text{base+DA}}$, highlighting the benefits of domain adaptation. Statistical tests also confirm that RoBERTa variants generally surpass ALBERT, BERT, and XLNet-based models in most metrics, except for precision, where the differences are not significant.

4.5.2 Assessing Models Generalizability and Robustness over Informative Samples with Varying Frequency Levels

Unseen or infrequent customer needs are crucial in gaining a competitive advantage in the market (Timoshenko & Hauser, 2019). Therefore, solely relying on the overall performance of a model may mask error patterns existing among informative samples with infrequent customer needs.

To assess models' generalizability across different groups of informative samples, we explore the relationship between the classification errors and the semantical appearance levels of informative samples during training. To achieve this, all the informative samples were sorted according to their similarity value, as defined in the section 4.3.2, and grouped into 12 clusters, each containing the same number of samples. Subsequently, the classification accuracy of the models was measured for each cluster. This experiment evaluates whether models can identify samples with infrequent customer needs as effectively as those with frequent customer needs, from a semantic perspective, thereby demonstrating their inter-domain generalizability. The results of this experiment are presented in Table 4.4.

Furthermore, evaluating the robustness of Transformer-based models when handling unseen informative samples presents another interesting avenue for exploring model performance. A dual-setting experiment was conducted in which either infrequent or highly frequent parts of informative samples were gradually excluded from the training phase and subsequently incorporated into the test set (i.e., unseen samples), which was later used to assess the robustness of the selected models (top-performing model within each family or variant). Since Transformer-based models have a known drawback of developing a bias towards frequently occurring vocabularies during training, this experiment further pinpoints the consequences of this tendency. The results of this experimental setup are depicted in Fig. 4.4.

Inter-domain Generalizability Assessment

Table 4.4 displays the classification accuracy for twelve clusters of informative samples grouped according to their similarity values. We consider informative samples with similarity values of less than 40 as semantically "*less-seen*" samples (clusters one and two) during training. Generally, all models failed to classify less-seen samples with the same level of accuracy as they exhibit for clusters with frequently seen samples, as confirmed by the performance gap in the table. This

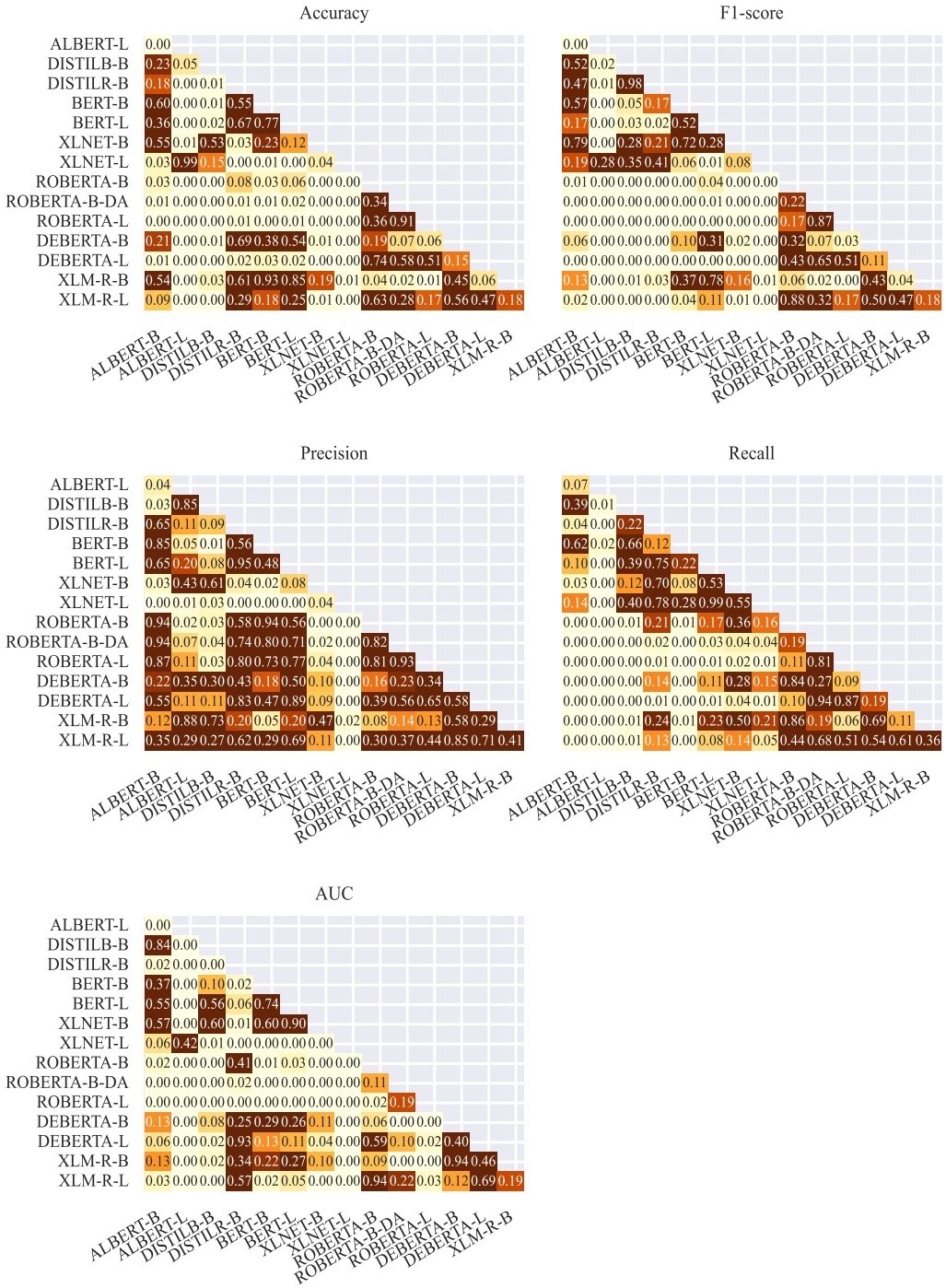

Figure 4.3: Statistical significance tests among different network architectures for Accuracy, F1-score, Precision, Recall, and AUC metrics. Adjusted P-values are shown in the cells. Light yellow indicates statistical significance with $p < 0.05$.

Table 4.4: Classification accuracy of Transformer-based models across twelve clusters of informative samples, grouped according to their similarity values from less-seen to highly seen samples. The table highlights significant performance discrepancies between the models on groups of semantically less-seen samples and more frequently observed samples during training, as demonstrated by the *MAX−MIN* column. The top performance across all models in each similarity-based cluster is in bold.

	Similarity-Based Clusters												
Distilled Models	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	**MAX−MIN**
DistilBERT	66.73	75.66	84.30	87.48	85.73	81.93	74.83	91.15	84.30	86.09	93.63	92.48	26.90
DistilRoBERTa	67.08	78.13	86.39	85.76	87.49	84.70	79.45	91.67	83.58	86.09	94.06	94.01	26.98
Base Models													
ALBERT	63.32	74.89	83.70	84.53	84.44	78.70	77.01	89.08	86.25	84.97	94.23	93.03	30.91
BERT	65.47	75.09	83.81	87.13	84.73	81.76	75.03	90.69	82.92	86.41	93.91	93.34	28.44
XLNet	68.97	77.93	87.28	87.34	86.89	85.61	77.15	90.92	84.44	87.50	94.66	93.89	25.69
RoBERTa	69.48	79.48	88.40	87.71	88.82	85.10	81.40	91.52	85.01	86.70	94.63	94.03	25.15
RoBERTa + DA	70.83	81.35	88.25	**90.00**	**90.23**	86.35	80.49	93.02	**88.00**	89.25	94.63	94.29	23.80
DEBERTa	69.51	81.26	88.88	86.65	87.78	85.78	81.08	90.95	85.13	88.71	94.60	94.01	25.09
XLM-RoBERTa	68.60	80.83	85.07	89.00	87.78	83.06	81.37	92.87	85.85	87.70	94.77	93.63	26.17
Large Models													
ALBERT	66.05	71.87	79.40	82.15	81.27	81.73	69.91	88.59	79.77	85.63	94.40	92.45	28.35
BERT	67.88	75.83	85.96	87.25	84.96	84.31	79.10	90.57	83.84	87.13	94.74	93.11	26.86
XLNet	70.14	78.53	81.35	87.11	85.16	87.28	76.31	91.44	84.04	88.19	93.91	91.27	23.77
RoBERTa	71.15	82.61	89.26	89.60	90.06	**88.67**	79.74	**93.25**	86.13	**89.80**	**95.17**	94.03	24.02
DEBERTa	**71.60**	**85.03**	**89.34**	89.20	89.08	88.24	**81.22**	91.75	84.15	88.53	94.80	**94.76**	23.20
XLM-RoBERTa	70.97	82.01	89.08	89.57	89.71	87.39	79.01	91.38	84.90	88.30	95.14	93.80	24.17
	Average Similarity Value of Samples in Each Cluster												
	14.42	37.71	54.23	75.28	126.52	150.71	161.21	201.97	243.46	283.32	428.67	509.38	

observation demonstrates that when these advanced models are not exposed to a sufficient number of similar versions of less-seen samples during training, they cannot fully learn generic patterns from the frequently seen parts to apply to less-seen samples. This deficiency leads to the misclassification of rare yet crucial samples.

The performance of models within the seventh cluster, which contains informative samples with an average similarity value of 161.21 during training, suggests that Transformer-based models can also have high classification errors over frequently observed samples. However, this trend does not necessarily stem from the models' inability to generalize. Rather, it underscores that even frequently mentioned sections of samples contain complex and contrastive patterns or even incorrect annotations, which complicate inter-domain generalization for models. Such samples necessitate cherry-pick analysis to identify common error patterns among them, an interesting pursuit that we reserve for future study.

Robustness Assessment on Unseen Informative Samples

This experiment bifurcates the original, informative data into remaining and excluded samples as dictated by a specified $Ratio$ criterion. In setting A, training is performed predominantly on highly frequent informative samples by gradually excluding informative samples from customer needs clusters with populations smaller than the $Ratio$ threshold, incorporating them into the *Excluded* test set. Conversely, setting B emphasizes training on infrequent informative samples by

gradually excluding the samples from the top-$Ratio$ most populated need clusters. Since informative samples might pertain to multiple need clusters, this introduces a semantic overlap between the training and testing sets. To eliminate semantic correlations between training and test datasets, we divided the *Excluded* test set into two strata: the *UnSeen* subset, containing unique samples not represented semantically in the training data, and the *Seen* subset, which includes samples that express customer needs already present in the training phase.

For both settings, we allocate 80% of the remaining data for training and 20% for the *Main* test set. Besides uninformative samples, in setting A, the *Main* test set gradually includes informative samples from less populated need clusters, while in setting B, it includes informative samples from highly populated need clusters. As the $Ratio$ value increases, the training sample size diminishes from 80% to around 10% of the total samples which would negatively affect the overall performance of models. Correspondingly, the amplification in the $Ratio$ expands the test set size, growing from 20% to 90% of the total samples. To preserve the data balance between both classes amid increasing $Ratio$ values, equal numbers of non-informative samples are randomly withdrawn for training in each experiment, facilitating a fair comparison.

Fig. 4.4 outlines the sensitivity analysis applied to the *UnSeen* and *Seen* subsets within the excluded, main, and aggregation of all test sets, along with the specificity analysis conducted on the main and aggregated test sets. Generally, all models exhibit significantly lower sensitivity performance on the *UnSeen* subset of the excluded test set compared to the *Seen* part of it. An interesting observation is that $\text{RoBERTa}_{\text{base+DA}}$ consistently outperformed both $\text{RoBERTa}_{\text{large}}$ and $\text{BERT}_{\text{large}}$ in classifying unseen informative samples across various settings. This finding emphasizes the potential of domain adaptation to enhance the robustness of Transformer-based models. Additionally, $\text{XLNet}_{\text{base}}$ also performed better than the mentioned large models, which shows the capability of this architecture in handling unseen samples.

From a bias analysis perspective, increasing the $Ratio$ in Setting A helps reduce the potential for lexical bias in models by excluding highly populated clusters (which often contain lexical cues). As shown in Fig. 4.4a, there is a sharp performance improvement up to the $Ratio$ of five. This implies that avoiding lexical bias enables models to learn more generic features that are advantageous when applied to unseen samples. Nonetheless, both the specificity and sensitivity of the main test set start to decline beyond the threshold of 11 in Setting A. This is anticipated since the complexity of the training data rises in this setting, and without lexical cues, models find it challenging to discern between informative and non-informative samples of the main test set.

On the other hand, increasing the $Ratio$ in Setting B intensifies the lexical bias, as the models are trained solely on data from the highly populated need clusters. Nonetheless, this does not significantly impair performance on unseen data up to the threshold of 190. This outcome indicates that, despite the heightened risk of lexical bias, Transformer networks maintained the task knowledge acquired during training to distinguish unseen informative samples from non-informative ones. Beyond this threshold, the models' sensitivity performance sharply declines, dropping below 20% on unseen data when training only included samples from the most populated need clusters. This performance drop is initially due to the overall decrease in the power of the model, given the fewer training samples used at each $Ratio$. More importantly, it arises from the very limited semantic diversity in the training samples, causing the models to overfit to the training data and rely heavily on lexical cues. This overfitting results in biased outputs that fail to reflect the true underlying patterns of the task and impairs their capacity to generalize to unseen data.

The above-mentioned worst-case scenario confirms the importance of utilizing an appropriate sampling method during the data collection, especially when data annotation resources are limited.

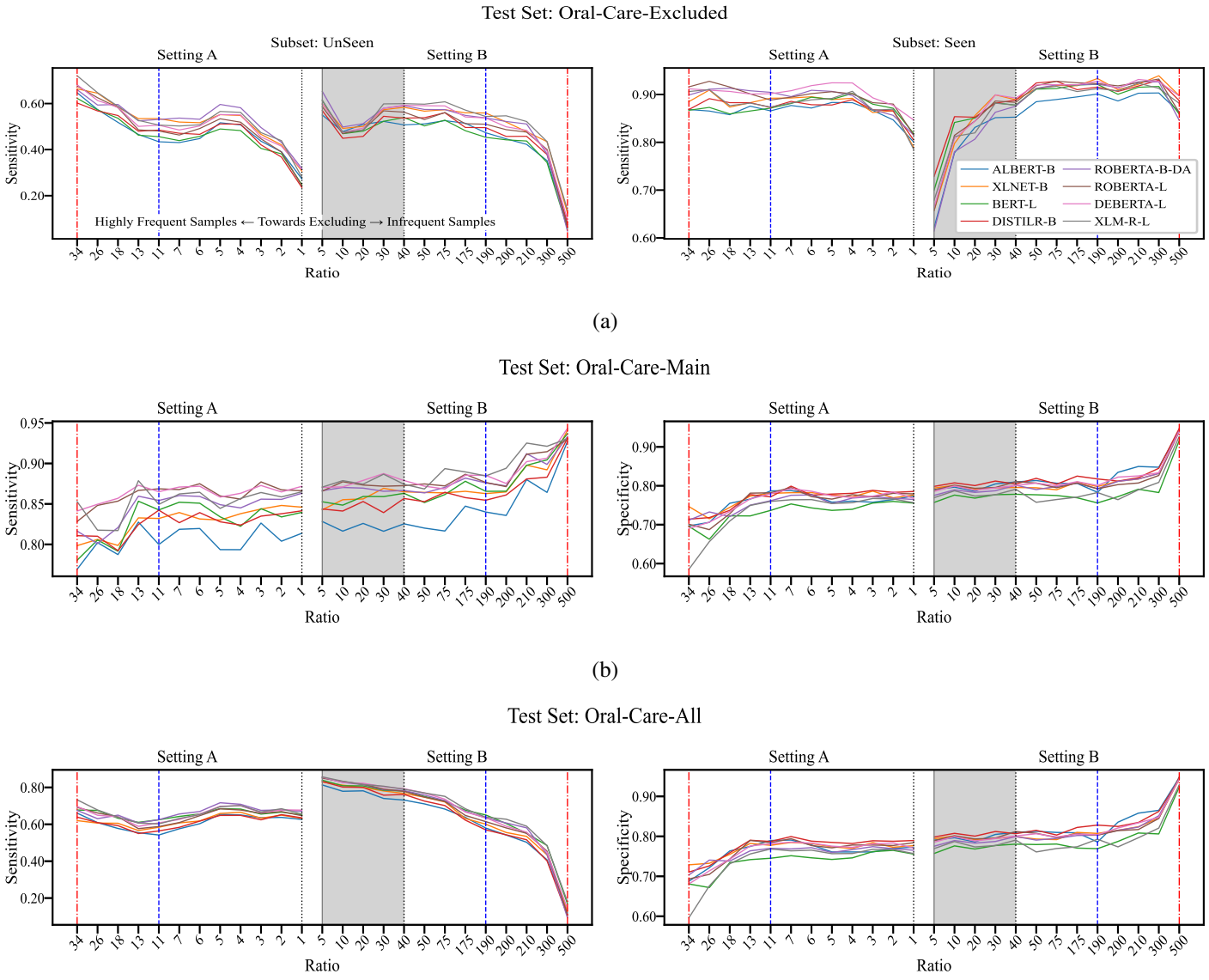

Figure 4.4: Robustness analysis of Transformer networks: In each figure, Setting A illustrates the results of models gradually trained on infrequent samples, whereas Setting B depicts a converse trend. (a) Shows sensitivity across unseen and seen parts of excluded test sets. An excluded sample is categorized as *Unseen* if it has no similar version in the training set (i.e., similarity value is zero); otherwise, it belongs to the *Seen* subset. (b) Highlights sensitivity and specificity on the main test set, while (c) displays sensitivity and specificity performance across an aggregation of all test sets. Vertical lines with matching patterns indicate comparison points between the two settings, showing roughly equal numbers of training samples.

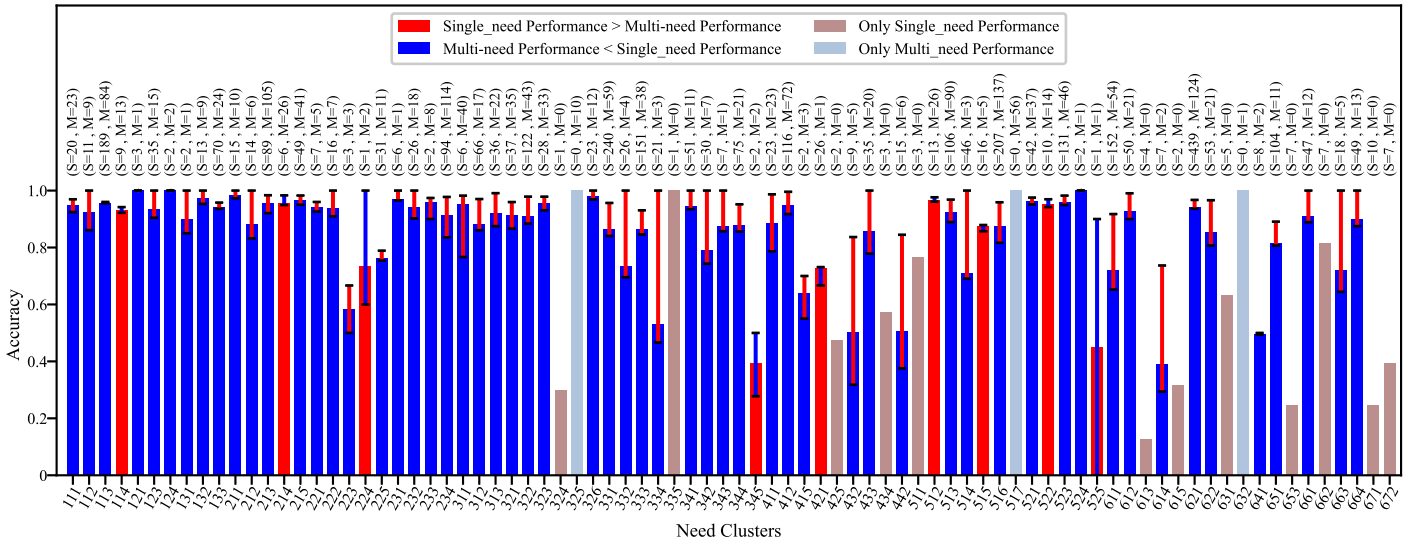

Figure 4.5: Average accuracy achieved by RoBERTa$_{\text{large}}$ on various need clusters within the informative class. Error bars represent the performance difference between the single-need and multi-need subgroups within each need cluster. Color-coded bars indicate the higher- and lower-performing subgroups and the presence or absence of subgroups in each cluster. The population of each subgroup is displayed at the top of each bar.

For instance, fundamental consumer needs, such as concerns about pricing, might be discussed more in reviews. After annotation, such samples could disproportionately represent most of the "informative" labels, thereby reducing the semantic diversity of the training data and introducing a lexical bias into the model.

Analyzing Model Performance on Need Clusters

In this experiment, each need cluster is divided into single- and multi-need subgroups, which helps to better understand the models' performance with less or more information in a sentence. Fig. 4.5 illustrates the average accuracy achieved by the best-performing model, RoBERTa$_{\text{large}}$, on all need clusters.

A close analysis of the results reveals significant differences in the accuracy scores between the two subgroups: the multi-need subgroup generally surpasses the single-need subgroup. This pattern suggests that Transformer-based network models also require a wealth of explicit information to make accurate decisions, supporting the lexical clue dependency finding discussed in the previous section. In particular, this trend generally holds regardless of the population size of the need clusters; even the large "611" cluster exhibits underperformance in single-topic samples. Despite that, the performance gap between single-topic and multi-topic samples is relatively minimal in some highly populated need clusters such as "621". A possible explanation for this nearly on-par performance is that these clusters contain lexical clues that have a stronger influence on the model's predictions.

To demonstrate the influence of lexical cues, we employed an error analysis method to generate two sets of tokens, each representing the top 20 tokens for each class that could affect the model's predictions. The tokens in each set were selected by sorting all unique tokens from the Oral-Care

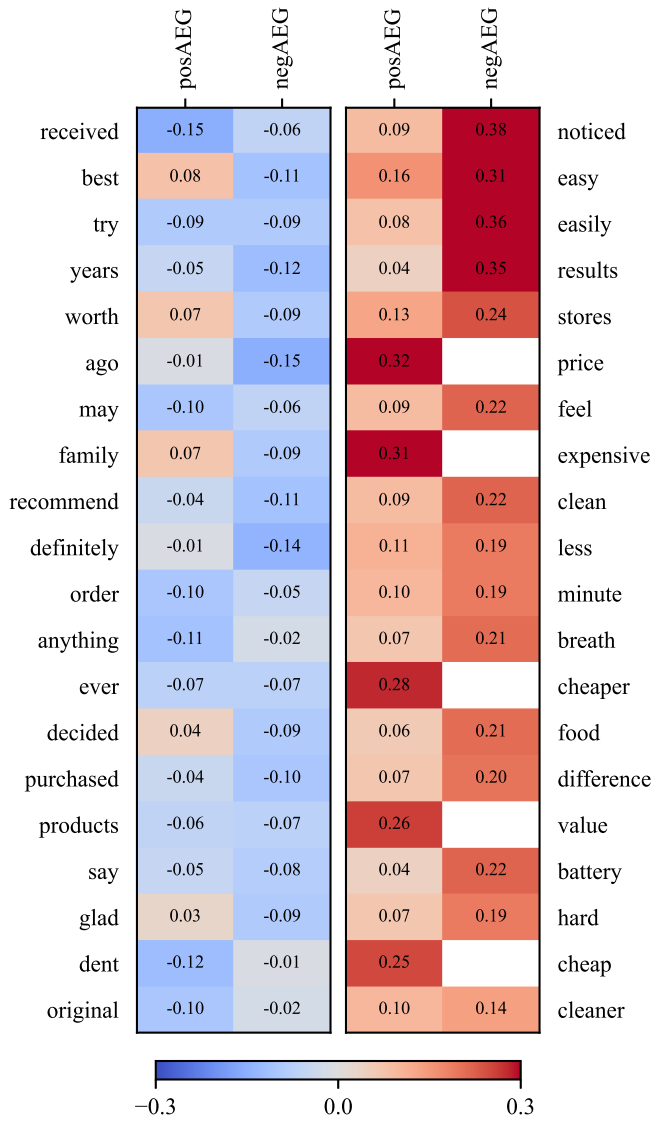

Figure 4.6: AEG values for the top 20 tokens in the informative (right) and non-informative (left) classes, ranked by their cumulative absolute AEG values. Selected tokens for the informative class appear in an average of 109.8 sentences (median 78.0) while sentences in the non-informative class have an average of 17.55 (median 11.0). Empty positions indicate that a token does not appear in any sample of that class.

dataset according to the sum of the absolute values of their AEGs. Utilizing fairness metrics in this context enables the estimation of the influence of specific tokens by comparing the prediction scores of samples containing a given token with those of all other samples within the same class. Fig. 4.6 illustrates RoBERTa$_{\text{large}}$ performance on both sets of tokens.

Fig. 4.6 shows that the model assigns significantly higher scores to sentences containing tokens from the right-set, thereby favoring the informative class. This can lead to an increased rate of false positives if they appear in non-informative samples, as indicated by their strong positive neg_{AEG}

value. The opposite trend is also true for the left-set tokens. A detailed analysis of the influence of both token sets on causing false predictions is given in Appendix C.

An additional observation is that the model is more sensitive to picking up lexical bias within the informative class than in the non-informative class. We presume this unintended intensification of bias toward the informative class can create complications. Specifically, if the model overrelies on lexical cues associated with the informative class, it may struggle to correctly identify informative samples that lack these cues, especially if those samples are infrequent which poses a significant challenge for this task. In future studies, we plan to delve deeper into the indirect influence of lexical cue bias on predicting rare, informative samples.

4.5.3 Sample Efficiency and Cross-Domain Adaptation

Sample Efficiency

In this section, we assess the sample efficiency of the selected models (top-performing model within each family or variant) using the Oral-Care dataset as the in-domain dataset and Electronics, Baby, Sports-Outdoors, and Pet Supplies datasets for out-of-domain evaluations, which are primarily imbalanced. Analyzing out-of-domain samples helps to understand to what extent models have learned features that are independent of any specific domain, and how the volume of training data affects their performance on data outside their training domain. The results of this analysis are presented in Fig. 4.7.

Considering the sizes of the models, larger models, with the exception of $\text{DeBERTa}_{\text{large}}$, underperformed in the in-domain setting when only a few examples were available for training, compared to the variants with smaller sizes. This result first shows that increasing the size of the models does not increase sample efficiency in the in-domain setting of this task, which is contrary to the finding in (N. F. Liu, Kumar, Liang, & Jia, 2022). Secondly, it illustrates the heightened risk of underfitting with larger models in the ICSN context for supervised learning when only limited samples are available. Moreover, while the in-domain performance generally improves with more training samples, the out-of-domain performance appears to plateau or decrease beyond the 10% ratio for all the models examined. This behavior indicates that out-of-domain performance does not linearly correlate with the number of in-domain training samples in this task. Nonetheless, maintaining a sufficient number of samples is necessary to achieve consistent out-of-domain results. Furthermore, this finding hints at the tendency of Transformer-based models to overly adapt to the training domain distribution beyond a particular data volume, losing the general knowledge acquired during pre-training. We presume this over-adaptation is the root cause limiting their generalization across the other four evaluated domains.

To further highlight the advantages of domain adaptation, we observed a positive impact on both the sample efficiency and out-of-domain performance of the $\text{RoBERTa}_{\text{base}}$ model. $\text{RoBERTa}_{\text{base+DA}}$ outperformed all other models in the limited-data training setting (using up to 10% of the training data) across the in-domain dataset and all out-of-domain datasets, except for the Electronics dataset.

Cross-domain Adaptation

To evaluate the domain adaptability of Transformer-based models, we conducted an experiment in which $\text{RoBERTa}_{\text{large}}$, was trained on each of the domains separately (source domain) and then assessed on all the remaining domains (target domains) without retraining on unseen domains.

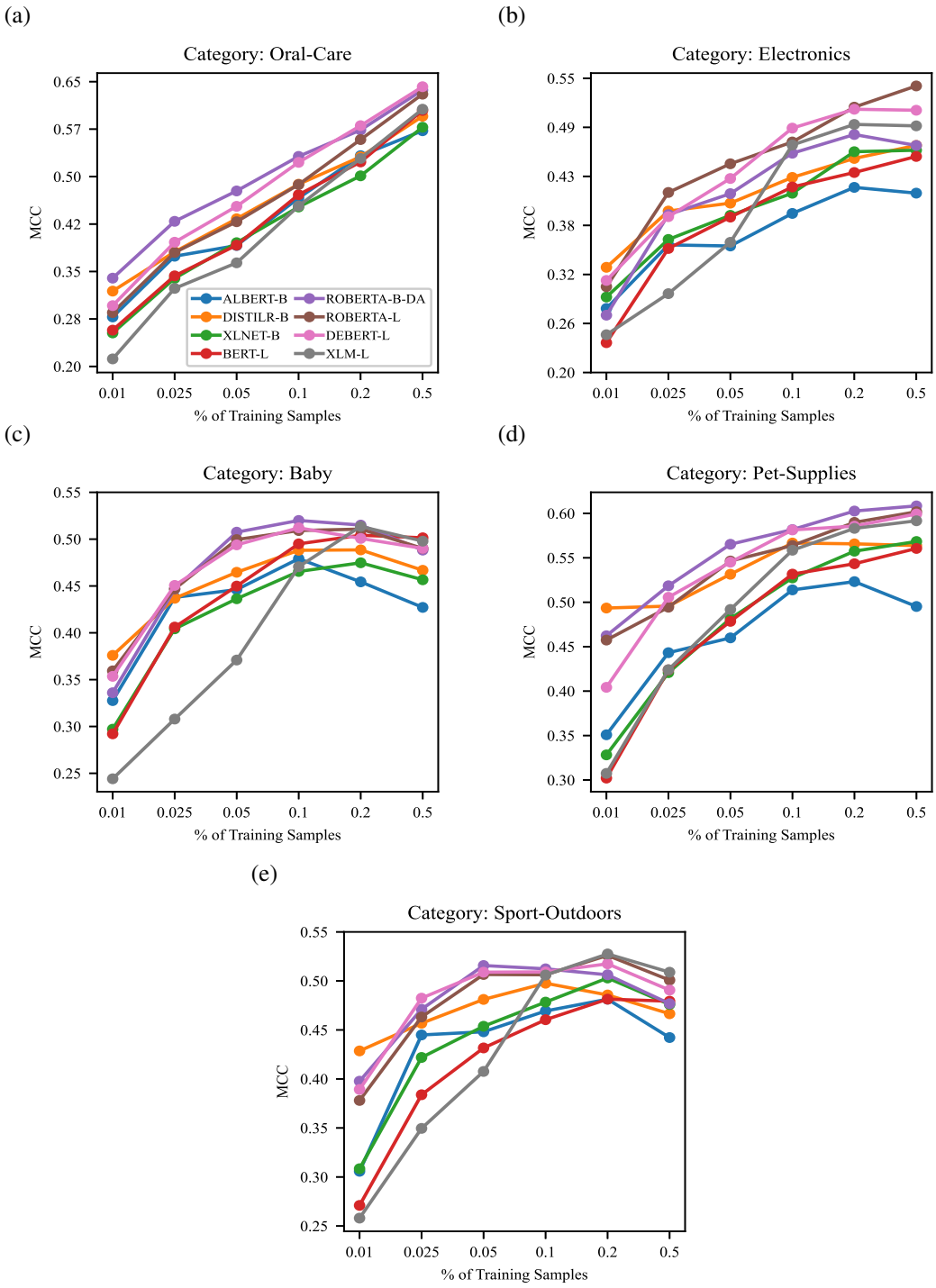

Figure 4.7: MCC results of in-domain (a) and out-of-domain evaluations in (b), (c), (d), and (e) for the sample efficiency experiment. While the in-domain performance shows improved outcomes with increased sample sizes, the out-of-domain performance exhibits a plateau or decreasing trend when utilizing more than 10% of the training samples.

Table 4.5: In-domain and cross-domain classification results of $\text{RoBERTa}_{\text{large}}$. Column averages indicate the model's generalization across target domains for each source domain, while row averages represent the prediction difficulty for each target domain across all source domains.

Target Domain	Metric	Source Domain					Mean ± STD
		Oral-Care	Electronics	Sports-Outdoors	Baby	Pet-supplies	
Oral-Care	MCC	55.1	39.2	37.5	32.2	37.6	36.6 ± 3.2
	TPR	83.2	61.7	62.2	56.0	50.5	57.5 ± 6.5
	TNR	71.3	77.1	75.1	75.5	85.0	78.2 ± 5.0
Electronics	MCC	51.8	60.6	46.2	50.5	34.0	45.6 ± 8.0
	TPR	73.7	81.4	57.6	61.8	25.4	54.5 ± 19.2
	TNR	78.5	80.2	86.1	86.5	97.2	87.0 ± 7.2
Sports-Outdoors	MCC	51.9	52.2	59.8	51.0	52.6	51.9 ± 1.8
	TPR	76.3	65.9	79.9	74.0	61.0	69.3 ± 7.1
	TNR	76.8	85.2	81.1	78.0	88.7	82.1 ± 5.4
Baby	MCC	50.9	58.8	57.7	63.6	53.4	55.1 ± 3.7
	TPR	84.8	75.6	80.0	84.5	66.3	76.6 ± 7.5
	TNR	66.2	83.0	78.1	79.6	85.6	78.2 ± 8.1
Pet-supplies	MCC	58.2	60.3	59.5	49.8	69.3	56.9 ± 4.9
	TPR	89.9	75.4	81.9	69.7	85.1	79.2 ± 9.3
	TNR	72.4	85.7	80.3	81.2	86.4	79.8 ± 5.6
MCC Mean ± STD		53.2 ± 3.7	52.6 ± 8.5	50.2 ± 9.2	45.9 ± 8.4	44.4 ± 9.2	–
TPR Mean ± STD		81.1 ± 7.6	69.6 ± 7.5	70.4 ± 11.8	65.4 ± 9.2	50.8 ± 16.5	–
TNR Mean ± STD		73.5 ± 6.1	82.7 ± 4.4	79.8 ± 4.8	80.3 ± 5.0	89.1 ± 5.1	–

Experiments were conducted solely on this model as it was identified in Section 4.5.1 as the best-performing model. The training was done with a consistent sample size of 1.6 thousand across all domains, achieved through 5-fold cross-validation. Since the Oral-Care dataset is four times larger than others, we used a reversed cross-validation strategy (four-fold for testing and one-fold for training) when using this dataset as the training source. This method maintained a uniform sample size across domains during training. Besides reporting MCC, sensitivity (i.e., TPR) and specificity (i.e., TNR) were also included in the evaluation. Results are reported in Table 4.5.

Training on Oral-Care datasets yielded the best performance in identifying informative sentences, achieving roughly 12% higher sensitivity than the second-best setting, which involves training on Electronics. Although the training data was nearly balanced, in-domain results from Oral-Care indicated a considerable skew towards the informative class, which resulted in the lowest specificity among of all settings. The high average sensitivity of Oral-Care across target domains may be attributed to the broad range of topics included in its informative class, given that the population of its informative samples is five times larger than that in the most balanced domain, Electronics, even though each model was trained using the same number of samples.

In contrast, when the training data is sourced from the other four domains, the model exhibits superior cross-domain specificity compared to sensitivity. This observation could be supported by the fact that these training sets contain more non-informative samples, as demonstrated in Table 4.2. Additionally, the best MCC result among these four datasets was obtained when using Electronics as the training data, which could be ascribed to the greater variety of unique tokens across both classes, presumably enhancing the model's ability to generalize and outperform in other domains.

Table 4.6: Separability Index values across various k-nearest neighbors (K).

Dataset	K_1	K_3	K_5	K_{10}
SST-2	84.1	80.7	79.2	77.4
IMDB	81.0	78.5	76.8	74.4
Rotten Tomatoes	75.7	74.8	74.4	73.9
Oral-Care	67.2	65.9	64.9	63.5
Electronics	72.5	70.7	69.9	67.9
Baby	72.6	72.2	72.0	70.8
Sports-Outdoors	74.6	72.8	72.3	71.0
Pet-supplies	79.6	78.9	78.1	76.5

Regarding the ease of predictability, models found the Oral-Care and Electronics domains challenging to predict, with average MCC scores of 36.6% and 45.6%, respectively, in contrast to the Sports-Outdoors, Baby, and Pet-Supplies categories, which were easier to predict.

To better understand what makes a dataset challenging to predict or an ideal choice as source domain data for the ISCN task, we examined the complexities of all datasets using the Separability Index value (Thornton, 2002) (see Appendix D). We hypothesize that training on more intricate domains can lead to a more transferable model that excels in simpler domains. We provide the SI values of each dataset in Table 4.6.

The table shows that Oral-Care and Electronics, which rank first and second in cross-domain performance, are the most complex domains regarding separability. Conversely, Pet-supplies is the most separable domain among ISCN datasets but fares the worst in cross-domain evaluation. This underscores the importance of the source domain's complexity when aiming for superior cross-domain performance beyond considering the evenness of the dataset and vocabulary diversity.

4.6 Limitations

This study presents certain limitations that need to be acknowledged. Access to well-defined clusters of informative samples in the Oral-Care dataset facilitated a thorough exploration of error patterns across these samples, considering their relative importance and frequency. While this was an ideal case, we believe an unsupervised topic modeling approach could be employed to cluster informative sentences, especially with the recent advancements in large language models. Although such clustering might not always be exact or easily interpretable, the literature suggests it remains effective for grouping similar samples, unveiling certain error patterns related to each group.

While the analysis of the Oral-Care dataset was exhaustive for informative samples, the oversight of non-informative samples might limit the comprehensive applicability of our conclusions. As previously mentioned, one can apply a topic modeling approach to reveal specific error patterns among non-informative samples, thereby gaining more comprehensive insights into the shortcomings of Transformer-based models in this task.

Moreover, this work is built on the assumption that test set labels are reliable across all domains of the ISCN task. The significance of this assumption cannot be understated, as it might not be valid in every scenario since annotation in the ISCN task can be very subjective. During the out-of-domain evaluation, consistent labeling was presumed for semantically identical samples across domains. However, given potential label variations for identical samples between domains, this can

complicate out-of-domain evaluations and skew reported performances.

Lastly, while the primary focus of this study was on the technical barriers to adopting ML solutions in customer needs analysis, the importance of analyzing the potential social implications of utilizing these models in real-world scenarios should not be overlooked. In future works, the authors plan to investigate the social impacts of deploying Transformer-based models in the context of the ISCN task.

4.7 Conclusion

Given the growing reliance of businesses on UGC to identify customer needs, the emphasis has been shifted from questioning the use of ML solutions to overcoming technical barriers to their adoption by ensuring robust application and efficacy through reliable evaluations. This study explored the efficacy of Transformer-based networks in the ISCN classification task by employing additional evaluation objectives tailored to the task, highlighting the overall performance of the models across different experimental settings and domains.

Our grouping analysis revealed that Transformer-based models exhibit similar performances on similarity-based clusters of informative samples irrespective of their size. Their predominant reliance on lexical indicators resulted in subpar performance on clusters containing infrequent samples, thereby exhibiting questionable robustness against unseen informative samples. A noteworthy observation was that when trained using just 5% of the total samples from a single customer needs cluster, the performance of the top-performing model, $\text{RoBERTa}_{\text{large}}$, declined drastically to below 10% for unseen informative samples. This underscores the importance of semantic diversity in training these models and highlights the need for effective sampling methods during data collection, especially when annotation resources are limited.

The models' reliance on lexical cues was further substantiated when analyzing highly populated need clusters. Using threshold-agnostic measures from the unintended bias analysis domain, we identified influential tokens causing incorrect predictions. Understanding these tokens can aid in developing bias mitigation strategies, which in turn can enhance the generalization and robustness of models in this task.

The sample efficiency experiments demonstrated that in-domain performance is not linearly correlated with the performance of models, as larger models struggled with only a few samples available. Moreover, cross-domain performance started to plateau, indicating a tendency to overadapt to the training domain beyond a certain threshold. Furthermore, upon examining the intra-domain generalizability of Transformer-based networks, we found that both domain complexity and imbalanced training data directly impact the cross-domain performance of the top model, $\text{RoBERTa}_{\text{large}}$.

For future research, we suggest focusing on improving the generalizability of Transformer-based models both within and across domains. There is a compelling need to strengthen model robustness against unseen samples and to develop strategies to pinpoint and reduce the effects of lexical bias in this field of study. While the previously discussed research paths address technical barriers to adopting ML for customer needs analysis, developing an evaluation system that comprehensively examines the societal implications of ML models remains essential. This step is vital to ensure the responsible and beneficial use of AI in real-world environments.

Chapter 5

Conclusion and Future Work

This thesis investigated intelligent approaches to identifying customer needs from UGC, with a particular focus on Transformer-based models, and emphasized the importance of treating this task as a distinct ML/NLP problem requiring comprehensive, taxonomy-aware, and fairness-oriented evaluation methods.

In Chapter 5, we discussed the key challenges of identifying customer needs from UGC and argued that they should be treated not simply as technical hurdles but as fundamental characteristics that define the task. The aim was to reframe common difficulties—such as inconsistent terminology, annotation subjectivity, data bias, temporal variation, and implicit needs—not as side issues, but as core elements that evaluation frameworks must explicitly address. This perspective highlighted the need for a more holistic approach to customer-needs identification, one that goes beyond accuracy on a single dataset and instead emphasizes transparency, robustness, and responsible use. To extend this discussion, we conducted a focused review of 35 recent works, categorizing their motivations and contributions. The review revealed a clear misalignment: while many studies recognize the difficulty of implicit or evolving needs, most concentrate on pipeline development or incremental improvements, leaving challenges such as identification of rare or unseen needs, development of interpretable automated methods,and responsible AI largely underexplored. The overall takeaway encourages researchers, particularly from computer science, to approach customer-needs identification as a distinct ML/NLP task that demands clearer constructs, taxonomy-aware evaluation, and explicit consideration of fairness and social context, rather than treating it as a generic application of text classification or topic modeling.

Building on the insights gained in Chapter 3, Chapter 4 focused on applying those lessons by developing a model for the ISCN task while considering a comprehensive evaluation framework aligned with the task's objectives and complexities. Particularly, the main purpose of evaluation framework was to better assess generalization and robustness, beside exploring the extent to which Transformer-based models are influenced by unintended bias. To this end, we introduced a series of evaluations that examined model performance from the mentioned perspectives, offering a more complete picture of both their strengths and limitations. The findings showed that although Transformer-based models outperform traditional machine learning and deep learning approaches by a considerable margin, their heavy reliance on lexical cues makes them highly sensitive to frequent tokens in the dataset. This tendency reduced precision and weakened their ability to recognize less frequent yet informative samples—often representing hidden opportunities in new product development—raising concerns about their robustness in real-world applications. More broadly, the results suggested that the success of these models cannot be judged by accuracy alone on a single

test set, but also depends on how well they handle linguistic diversity in customer needs and resist overfitting to surface-level patterns. Taken together, the outcomes of this chapter reinforced the importance of evaluating models beyond simple performance scores and motivated us to extend our investigation into a new context of social bias analysis of these methods which we leave for future work.

In summary, this thesis shows that evaluating customer-needs identification from multiple perspectives is essential for understanding model strengths and limitations. The findings guide future research toward improving robustness, fairness, and interpretability while bridging gaps between marketing perspectives and ML/NLP methods. Taken together, the work positions customer-needs identification as a distinct research area with opportunities for both technical progress and practical impact.

5.1 Limitations

This section explains the limitations of the thesis and points out possible challenges and biases that may have affected the results.

In Chapter 3, the review was conducted using a non-exhaustive search strategy, which means some relevant studies may have been overlooked. In addition, the categorization of motivations and contributions relied on our own interpretation, introducing subjectivity that may have influenced the conclusions drawn.

In Chapter 4, the analysis relied on well-defined clusters of informative samples within a specific dataset, which may limit how broadly the findings apply to other domains or less structured datasets. Moreover, the evaluation assumed that labels across domains were consistent and reliable, but given the subjectivity of the ISCN task, this assumption may not always hold, potentially affecting the validity of out-of-domain results.

5.2 Directions for Future Research

Future work should first expand the Chapter 3 review into a full systematic literature review that widens coverage beyond need identification and uses protocol-driven methods to examine, at scale, how motivations, contributions, and limitations align across methodologies.This would reduce selection bias, improve reproducibility, and provide a more rigorous understanding of the misalignment between motivations, contributions, and limitations observed in existing studies.

Second, the work in Chapter 4 can be extended by focusing on improving model generalization within and across domains, enhancing robustness to rare or unseen samples, and reducing reliance on frequent lexical cues. Another interesting avenue for future work would be to investigate whether the observed bias toward frequent lexical cues in the informative class directly affects the detection of rare but informative samples, and to examine the extent to which this influence shapes overall model performance.

Third, Chapter 5 points to the need for clear definitions and practical metrics for social bias in customer-needs pipelines, including fully automated systems that employ generative AI, whether proprietary or open-weight, in order to understand how bias arises, how it impacts the outcomes of such systems despite pre-release mitigation efforts, and who is affected by these outcomes.

Finally, a general recommendation is for researchers in this area to organize workshops aimed

at improving standards in the field, particularly by clarifying task definitions and introducing benchmarks and datasets that can serve as common points of comparison, which the field currently lacks.

Appendix A

Unsupervised Annotation

In this section, we present a preliminary assessment of the annotating capabilities of AI-based tools. The objective is not only to evaluate their functionality but also to explore the potential benefits of using these tools, such as cost-effectiveness, improved efficiency, and scalability. Our tool of choice for this annotation task is ChatGPT 3.5 turbo (OpenAI, 2023), which was deployed to unsupervisedly annotate datasets as introduced in Section 4.5.1 with the presumption of accepting their labels as the "gold standard", thereby comparing the annotation ability of ChatGPT 3.5 turbo with the existing labels.

To carry out the annotation process, two prompts per sentence were created using the customer need definitions outlined in Section 3.2.1, with a restrained application of prompt engineering techniques solely to yield parsable output. For each prompt, we explored two temperature settings, zero and one, to modulate the determinism of the model's output. Consequently, four ChatGPT responses were generated for each sentence, totaling 24,000 requests for the 8,000 samples included in each dataset, at the cost of $22.18 for both datasets, covering approximately 11 million tokens. It is important to note that the API environment of ChatGPT does not carry over the annotation history, as each request initiates a new chat history. The zero-shot model's accuracy is illustrated in Figure A.1.

In terms of accuracy, prompts derived from the customer need definition outlined in (Timoshenko & Hauser, 2019) outperformed those generated from the second definition, despite both sets achieving over 60% agreement with the golden labels. This is promising, as the model's performance is at least 10% above random guess performance, and its capability can be dramatically increased if appropriate prompts are provided. While we did not have access to OpenAI's latest chat model, GPT-4, which is more powerful than the version we used, we believe this model could potentially enhance performance to an acceptable range. Furthermore, model accuracy remained unaffected by different temperature values, despite occasional inconsistencies when handling non-coherent sentences such as those comprising a single word in our datasets. Yet, after enhancing prompts by modifying action requests, these inconsistencies were successfully addressed in an additional investigation, resulting in relevant model outputs. The intercoder agreement was not evaluated in our study because it did not involve prompt engineering techniques, and the two prompts used were distinct (different definitions).

As for scalability and efficiency, and considering OpenAI's API restrictions (OpenAI, 2023b), all annotation processes were completed in under 40 hours, with an average of 1,200 API requests per hour. This efficiency surpasses that of human annotators, which we evaluated in a small-sample study with three research assistants. Based on our findings, each annotator was able to annotate

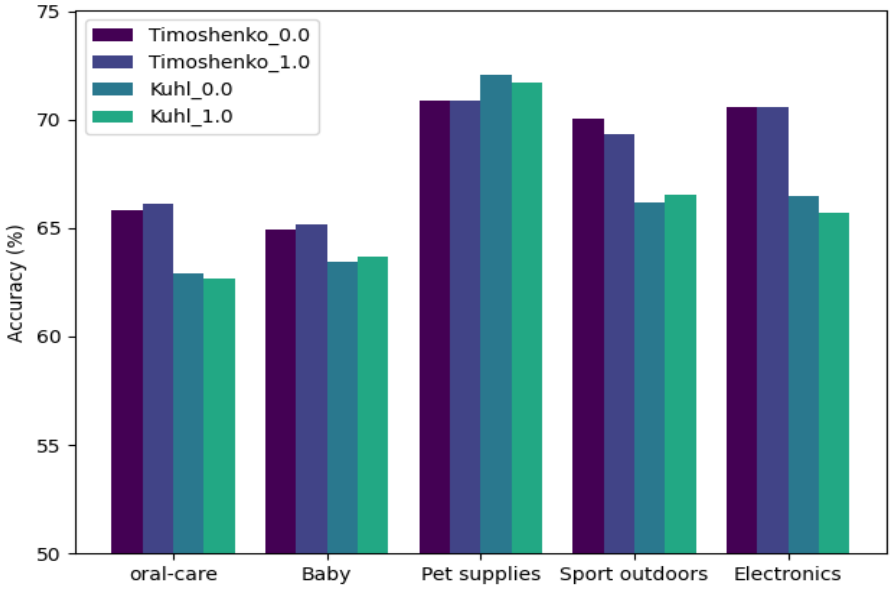

Figure A.1: Performance of ChatGPT in zero-shot text annotation, measured by accuracy in agreement with golden labels.

approximately 200-300 samples per hour, which is significantly lower than ChatGPT performance since its API limit can also be increased upon user request.

Despite ChatGPT's encouraging performance, one must not ignore potential inconsistencies and reliability concerns, given that varying responses can be generated for identical requests. Therefore, it remains crucial to establish comprehensive validation measures to ensure the model's reliability. In this context, an intriguing approach can be the use of an aggregation strategy Gilardi et al. (2023); Reiss (2023) in which the model's reliability and consistency can be significantly improved by pooling the outputs from repeated iterations of the same and different prompts. The implementation of such a method could further enhance the overall performance and reliability of ChatGPT, thus making it more useful for practical applications.

Appendix B

Structured Review Results

Table B.1 present summary of over idea of selected papers in the structure review study discussed in Section 3.4.

Table B.1: General overview of the reviewed literatures

Reference	Terminology	General Idea
Jhamtani et al. (2015)	✓	– The authors proposed a system that uses linguistic patterns and supervised machine learning to automatically identify product improvement suggestions in online product reviews. Their approach combines natural language processing with classification models to distinguish suggestions from other review content.
Kuehl et al. (2016)	✓	– The authors propose a Machine Learning approach to identify micro blog posts that express customer needs. Following a Design Science Research framework, they developed a classification artifact through a five-step process: data retrieval (via Twitter API and IBM Insights), data coding (manual descriptive coding), data filtering (language, URLs, duplicates), data labeling (crowdsourced classification), and supervised learning (preprocessing, sampling, and testing various algorithms). This enables scalable elicitation of customer needs from large datasets.
Guzman et al. (2017)	✓	– The authors present ALERTme, an approach to automatically classify, group, and rank tweets about software applications. They use supervised machine learning for classification, Biterm Topic Modeling for grouping related tweets, and a weighted ranking function to prioritize tweets according to attributes such as sentiment, category, likes, retweets, and duplicates.
(C. Li et al., 2018)	✗	– The authors propose a keywords-based machine learning approach to semi-automatically classify user requests in crowdsourcing requirements engineering. It combines non-project-specific and project-specific keywords, heuristic properties of user requests, and active learning strategy, with classifiers built using k-NN, Naïve Bayes, and SVM.

Continued on next page

Reference	Terminology	General Idea
Timoshenko and Hauser (2019)	✓	– The authors propose a hybrid method using machine learning to identify customer needs from user-generated content (UGC). Their five-stage process involves preprocessing UGC, training word embeddings, applying a CNN to filter non-informative content, clustering sentence embeddings to reduce redundancy, and having professional analysts manually extract customer needs.
Ayoub et al. (2019)	✗	– The authors propose a machine-learning approach to analyze customer needs in product ecosystems by filtering uninformative reviews with fastText, extracting topics using latent Dirichlet allocation (LDA), predicting sentiment and intensity with VADER, and categorizing needs with an analytical Kano model
M. Li et al. (2020)	✗	– The author proposes using multi-task learning (MTL) with hard parameter sharing to jointly address the tasks of requirements discovery (RD) and requirements annotation (RA). In this approach, requirements discovery (RD) is framed as a binary classification problem, where the goal is to determine whether a new document qualifies as a valid requirement. Meanwhile, requirements annotation (RA) is treated as a multi-label classification problem, focusing on assigning semantic categories to the sentences within the document
Kocon et al. (2021)	✗	– The authors propose the Automatic Aspect-Based Sentiment Analysis (AABSA) model, which automatically identifies hierarchical aspects (hypernyms and hyponyms) from Chinese online reviews using k-means clustering, BERT-based sentence embeddings, word2vec, and PageRank, and then performs sentiment analysis with Maximum Entropy (MaxEnt). The model is fully automated, domain-knowledge agnostic, and applied to Alibaba product reviews.
Kovacs et al. (2021)	✓	– The authors propose an unsupervised approach to assess customer needs from online reviews by combining topic modeling (LDA) with linguistic cues, followed by semantic consistency ranking using ELMo embeddings, and a manual aspect revision step. The method extracts, ranks, and refines product/service aspects from nearly 64 million Japanese customer reviews to generate requirement candidates.
De Araújo and Marcacini (2021)	✗	– The authors propose RE-BERT (Requirements Engineering using Bidirectional Encoder Representations from Transformers), which uses pre-trained BERT language models fine-tuned with a focus on local and global contexts for token classification in app reviews, enabling automatic extraction of software requirements.
Han and Moghaddam (2021)	✓	– The authors propose new formulation for CN extraction (ACOSI) with five labels (aspect, category, opinion, sentiment, implicit indicator) and develop a unified deep learning–based NLP model (fine-tuned T5) that extracts all labels simultaneously in a generative manner. The framework automates large-scale elicitation of implicit user needs from online product reviews.

Continued on next page

Reference	Terminology	General Idea
M. Zhang et al. (2021)	✓	– The authors propose a deep learning-based approach (REE-LSTM) to identify sentences in online reviews that contain innovation ideas. They develop a novel RNN-based Ensemble Embedding (REE) method combining GloVe, BERT, and XLNet embeddings, use a bidirectional LSTM for classification, and incorporate a focal loss function to address class imbalance.
Bian et al. (2022)	✗	– The authors develop a refined fine-grained sentiment analysis methodology with four steps: sentiment element extraction (Bi-LSTM+CRF), aspect-opinion pair identification (improved CNN combining structured and unstructured features), sentiment value calculation (dictionary-based with modifiers), and aspect term clustering (word2vec + K-means) to identify customer preferences from hotel online reviews.
Q. Zhao et al. (2022)	✗	– This method is divided into three parts. Firstly, text mining is adopted to collect online review data of multi-generation products and identify product attributes. Secondly, the attention and sentiment scores of product attributes are calculated with a natural language processing tool, and further integrated into the corresponding satisfaction scores. Finally, the improvement direction for next-generation products is determined based on the changing satisfaction scores of multi-generation product attributes.
Xiao et al. (2022)	✗	– The paper proposes a user preference mining method based on fine-grained sentiment analysis, modeled as a sequence labeling problem. It integrates a pre-trained BERT model to encode contextual user features, incorporates linguistic knowledge (POS and segmentation), and applies multi-scale convolution to capture text features at different scales. Finally, a Conditional Random Field (CRF) decodes the optimal label sequence, enabling accurate sentiment polarity detection in user reviews.
Mahdi et al. (2022)	✗	– The authors propose an Idea Mining framework with three stages: (1) Filtering – a classifier (feedforward neural network with BERT encoder and preprocessing layer) removes non-suggestive reviews, (2) Similarity Measures – cosine similarity with BERT-based sentence embeddings clusters similar suggestive reviews, and (3) Evaluation – assessing ideas as good/bad using NLP and statistical factors.
Salminen et al. (2022)	✗	– The authors collected 4.2 million tweets targeting 20 global brands from five industries, annotated samples, and trained multiple machine learning (ML) models to detect customer pain points and their types. They compared algorithms, optimized neural networks and transformer-based models, and proposed "pain point profiling" to categorize issues into five classes for managerial insights.
Stahlmann et al. (2023)	✓	– The study benchmarks previously proposed needmining models (SVM, naïve Bayes, CNN, RNN ensemble, RoBERTa) using a newly created publicly available gold set of annotated Amazon product reviews. They use design science research, build the dataset, validate inter-rater agreement, label sentences, and evaluate models with cross-validation.

Continued on next page

Reference	Terminology	General Idea
Cong et al. (2023)	✗	– The paper proposes a small sample data-driven method (ERNIE-ISIFRank) for eliciting user needs from online reviews. The framework has two stages: (1) topic-based classification of online reviews using ERNIE; (2) extraction of key product information phrases (PIPs) with improved SIFRank (ISIFRank), then manual transformation into explicit user needs.
Lee et al. (2023)	✗	– The authors propose a context-aware approach using linguistic pattern mining on online product reviews. The method extracts context information and product functions, clusters them using word embedding and k-means, and identifies customer needs by analyzing co-occurrence of context and function clusters.
Kaur and Kaur (2023)	✗	– The paper proposes MNoR-BERT, a transfer learning-based framework using BERT to classify multi-label user reviews into non-functional requirements (NFRs). It evaluates performance on a dataset of 6000 app store reviews, comparing with baseline machine learning, deep learning, and keyword-based approaches.
M. Zhang et al. (2023)	✗	– The authors propose a framework combining initial and supplementary online reviews to identify dynamic customer requirements. They use LDA to extract product attributes, machine learning–based sentiment analysis for aspect orientation, SnowNLP for overall satisfaction, regression to measure attribute effects, and the Kano model to classify product attributes.
Han et al. (2023)	✓	– The study proposes a context-aware approach for identifying customer needs from online product reviews. It extracts context information and product functions using linguistic pattern mining, clusters them with BERT and k-means, then defines customer needs through co-occurrence analysis of context and product function clusters.
K. Zhang et al. (2023)	✓	– The authors develops the UNISON framework, a systematic, data-driven approach for eliciting and evaluating smart product-service system (PSS) requirements. Using user online reviews as data, they apply Bi-LSTM for classifying product- and service-related requirements, and BTM topic modeling for requirement elicitation. For evaluation, they integrate sentiment analysis, IPA-Kano model, and the opportunity algorithm to assess, classify, and prioritize requirements. An empirical study on smart cleaning robots validate the framework's effectiveness.
Q. Li et al. (2023)	✗	– The study proposes a user demand mining method that integrates online reviews and complaint information. Product attributes are extracted using TF–IDF, expert consultation, and LDA; aspect-level sentiment is analyzed with a fine-tuned BERT model; complaint information is manually labeled and classified using BERT. Results from sentiment analysis and complaint classification are integrated to obtain comprehensive user demand elements.

Continued on next page

Reference	Terminology	General Idea
Yin et al. (2023)	✗	– The authors develop a framework based on lead user theory and machine-learning algorithms to automatically capture improvement ideas from social media chatter. The framework includes data preprocessing, improvement chatter identification using features of lead users and text, imbalanced classification methods, and topic-modeling-based summarization for managers
Z. Zhang et al. (2024)	✗	– The framework combines ERNIE 3.0-LDA-K-Means for topic modeling of product attributes, fine-tuned SKEP for sentiment analysis, and the Kano model for requirement classification, using online product reviews (e.g., smartphones) to improve accuracy and reduce dependence on large annotated datasets.
C. Wang et al. (2024)	✓	– The authors propose a customer needs mining framework using LLM agents to transform unstructured user-generated content into structured customer needs. The process involves classification, alignment, feasibility analysis, product improvement, and expert analysis, with LLM agents performing specific roles through prompt design.
Han and Moghaddam (2024)	✗	– The paper proposes the ACOSI (Aspect, Category, Opinion, Sentiment, Implicit indicator) analysis task and develops a unified model based on T5 transformers. It introduces a Design-Knowledge-Guided (DKG) position encoding algorithm and a domain knowledge benchmark (DKG-ROUGE) to extract implicit knowledge from online product reviews in a generative manner.
Barandoni et al. (2024)	✓	– The authors conduct a comparative analysis of various open-source and proprietary LLMs (e.g., GPT-4, Gemini, Mistral) to extract travel customer needs from TripAdvisor posts. They manually label needs, designed prompts (few-shot and optimized Chain-of-Thought via DSPy), deploy models, and evaluate outputs using BERTScore, ROUGE, and BLEU against the manual annotations.
Ettrich et al. (2024)	✓	– The study reconceptualizes Needmining from a binary classification task to a token classification task. The authors develop an artifact that identifies attributes and characteristics in user-generated content using transformer-based models and token classification, thereby extracting specific customer needs and organizing them to support decision making.
Kilroy et al. (2024)	✓	– The study builds a supervised Multivariate Time Series Classification (MTSC) model trained on the Trending Customer Needs (TCN) dataset and Reddit posts. It incorporates Multi-Task Learning (MTL) across multiple product categories, enabling prediction of future customer needs (1–3 years ahead) even in categories unseen during training.
Huang et al. (2025)	✗	– The authors reframe identifying novel customer needs as a text classification problem using a regularized dual BERT structure to analyze online product reviews. They mitigate class imbalance by introducing Kullback-Leibler (KL) divergence as a regularization mechanism, enabling robust and accurate detection of novel needs from user-generated content.

Continued on next page

Reference	Terminology	General Idea
Timoshenko et al. (2025)	✓	– The authors evaluate whether large language models can extract customer needs from qualitative data. They compare three approaches: (1) Base LLM with prompt engineering, (2) supervised fine-tuned (SFT) LLM trained on professional CN data, and (3) professional analysts, using blind studies across multiple product categories.
W. Wei et al. (2025)	✗	– The authors propose a framework based on large language models (LLMs) to analyze user needs from user-generated content (UGC). The method involves four steps: (1) collecting and preprocessing UGC data with a self-developed crawler, (2) extracting product attributes using LLMs and normalization, (3) conducting sentiment analysis via LLM embeddings and multilayer perceptron classification, and (4) mapping attributes into a quantified IPA-Kano model to prioritize product features and support user-centric optimization strategies.

Appendix C

The Role of Lexical Cues in False Predictions

We presented precision scores for the top 20, 35, and 50 tokens identified by our method as prominent lexical clues for all types of prediction outcomes in Table C.1. This table shows the contribution of lexical bias to the model's performance. Specifically, false positives contain more lexical clues that cause right score shift (RSS) than lexical clues that cause a left score shift (LSS). This pattern is also observed in single-topic false negatives that contain more LSS tokens, generally associated with non-informative classes, than RSS tokens. For example, only 33% of false negatives in the infrequent single-topic informative subgroup contained the top 50 RSS tokens. In contrast, 64% of them had LSS tokens, which makes these informative samples difficult for the model to classify correctly. In comparison, multi-topic false negatives contain slightly more RSS tokens than LSS ones, which complicates the conclusion that LSS-inducing tokens are the primary contributors to their misclassification. Our analysis also shows that the precision score increases as the token set size expands across informative and non-informative samples. This indicates our method's efficacy in accurately identifying the most critical tokens from both classes.

In summary, false positives are largely affected by the presence of RSS tokens. Similarly, false negatives, which are the primary concern in this application, are mainly triggered by left score shift tokens and are typically found among single-topic samples. Our findings emphasize that identifying and mitigating lexical bias is pivotal in improving the generalization, robustness, and fairness of Transformer-based networks, especially when they are utilized in challenging tasks such as ISCN.

Table C.1: Precision scores for prominent lexical cues in informative and non-informative samples, categorized by frequency and type (single-topic and multi-topic). RSS denotes right-side tokens displayed in Fig. 4.6, which contribute to a Right Score Shift in predictions, whereas LSS represents left-side tokens from the same figure, contributing to a Left Score Shift.

Rank	Category	Subgroup	Metric	Single Topic			Multi Topic		
				RSS	LSS	Shared	RSS	LSS	Shared
Top 20	Non-informative	All Samples	TN	5.09	26.88	1.09			
			FP	20.36	16.30	3.58			
	Informative	Infrequent	TP	29.31	13.89	5.56	50.37	18.41	11.34
			FN	10.75	23.53	2.54	32.08	27.36	16.04
		High FR.	TP	29.31	15.02	4.92	53.94	16.29	9.29
			FN	14.89	19.94	4.68	35.79	46.32	30.53
		Very High FR.	TP	44.91	17.18	8.40	64.59	16.92	11.19
			FN	15.44	19.68	2.19	47.93	44.24	34.10
Top 35	Non-informative	All Samples	TN	16.05	43.11	5.65			
			FP	44.10	29.65	13.20			
	Informative	Infrequent	TP	57.58	27.63	16.10	78.64	30.57	25.23
			FN	21.10	42.19	11.66	50.00	35.85	24.53
		High FR.	TP	57.47	29.47	16.23	82.55	29.49	24.78
			FN	37.98	37.88	16.22	50.53	51.58	36.84
		Very High FR.	TP	71.01	31.25	22.77	90.04	31.66	28.60
			FN	38.94	38.55	16.40	69.12	56.68	51.15
Top 50	Non-informative	All Samples	TN	29.58	63.68	16.74			
			FP	50.71	50.02	25.73			
	Informative	Infrequent	TP	70.65	46.09	30.80	92.28	43.30	38.21
			FN	33.67	64.71	24.75	59.43	56.60	42.45
		High FR.	TP	71.81	43.69	30.12	92.86	44.40	41.33
			FN	51.34	51.91	24.90	68.42	62.11	54.74
		Very High FR.	TP	79.43	48.75	39.57	93.96	48.76	45.36
			FN	49.93	57.17	28.06	76.50	76.50	62.67

Appendix D

Separability Index Value

The separability Index value is calculated for different k-nearest neighbor settings based on Euclidean distance, revealing the complexity of the task. The $SI(i,k)$ denotes the normalized proportion of i's k-nearest neighbors from the same class as sample i. The generalized formula is given by:

$$SI(i,k) = \frac{1}{k} \cdot |\{j \in S_k(i) \mid c(i) = c(j)\}| \tag{15}$$

where, $S_k(i)$ is the set of i's k-nearest neighbors, and $c(x)$ signifies the class of x.

References

Abbas, Q. (2024). The impact of personalization strategies on consumer engagement and conversion rates in digital marketing. *International Journal of Advanced Multidisciplinary Research and Study*, *4*(1), 452–454.

Agarwal, O., & Nenkova, A. (2022). Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, *10*, 904–921. doi: 10.1162/tacl_a_00497

Aldunate, Á., Maldonado, S., Vairetti, C., & Armelini, G. (2022). Understanding customer satisfaction via deep learning and natural language processing. *Expert Systems with Applications*, *209*, 118309. doi: 10.1016/j.eswa.2022.118309

Almagrabi, H., Malibari, A., & McNaught, J. (2018). Corpus analysis and annotation for helpful sentences in product. *Computer and Information Science*, *11*(2), 76. doi: 10.5539/cis.v11n2p76

Al Nefaie, M., & Muthaly, S. (2022, July). An effective hybrid data analytics technique for a 360-degree view of customer data. In *Proceedings of the 16th International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing (CGVCVIP 2022), the 8th International Conference on Connected Smart Cities (CSC 2022), 7th International Conference on Big Data Analytics, Data Mining and Computational Intelligence (bigdaci'22) and 11th International Conference on Theory and Practice in Modern Computing (TPMC 2022).* IADIS Press. doi: 10.33965/MCCSIS2022_202206C029

An, H., Acquaye, C., Wang, C., Li, Z., & Rudinger, R. (2024, August). Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers)* (pp. 386–397). Bangkok, Thailand: Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.37

Ana, M.-I., & Istudor, L.-G. (2019, March). The role of social media and user-generated-content in millennials' travel behavior. *Management Dynamics in the Knowledge Economy*, *7*(1), 87–104. doi: 10.25019/mdke/7.1.05

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of Data and Analytics* (pp. 254–264). Auerbach Publications.

Arora, C., Grundy, J., & Abdelrazek, M. (2024). Advancing requirements engineering through generative AI: Assessing the role of llms. In A. Nguyen-Duc, P. Abrahamsson, & F. Khomh (Eds.), *Generative AI for Effective Software Development* (pp. 129–148). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-55642-5_6

Ataei, M., Cheong, H., Grandi, D., Wang, Y., Morris, N., & Tessier, A. (2024). *Elicitron: A framework for simulating design requirements elicitation using large language model agents* (Vol. Volume 3B: 50th Design Automation Conference (DAC)). doi: 10.1115/DETC2024

-143598

Ayoub, J., Zhou, F., Xu, Q., & Yang, J. (2019, August). Analyzing customer needs of product ecosystems using online product reviews. In *Volume 2A: 45th Design Automation Conference* (p. V02AT03A002). Anaheim, California, USA: American Society of Mechanical Engineers. doi: 10.1115/DETC2019-97642

Bansal, P., & Sharma, A. (2023). Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *Arxiv Preprint Arxiv:2306.15766*.

Bao, Y., Wei, Z., & Di Benedetto, A. (2020). Identifying the tacit entrepreneurial opportunity of latent customer needs in an emerging economy: The effects of experiential market learning versus vicarious market learning. *Strategic Entrepreneurship Journal*, *14*(3), 444–469. doi: 10.1002/sej.1350

Barandoni, S., Chiarello, F., Cascone, L., Marrale, E., & Puccio, S. (2024). *Automating customer needs analysis: A comparative study of large language models in the travel industry.* arXiv. doi: 10.48550/ARXIV.2404.17975

Barnham, C. (2015). Quantitative and qualitative research: Perceptual foundations. *International Journal of Market Research*, *57*(6), 837–854. doi: 10.2501/IJMR-2015-070

Barocas, S., Crawford, K., Shapiro, A., & Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of SIGCIS.* Philadelphia, PA.

Bartl, M., Nissim, M., & Gatt, A. (2020, December). Unmasking contextual stereotypes: Measuring and mitigating BERT‘s gender bias. In M. R. Costa-jussà, C. Hardmeier, W. Radford, & K. Webster (Eds.), *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing* (pp. 1–16). Barcelona, Spain (Online): Association for Computational Linguistics.

Baxter, J. (2000, March). A model of inductive bias learning. , *12*(1), 149–198.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. doi: 10.1109/TPAMI.2013.50

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003, March). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(null), 1137–1155.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the tribes: Using text for marketing insight. *Journal of Marketing*, *84*(1), 1–25. doi: 10.1177/0022242919873106

Bian, Y., Ye, R., Zhang, J., & Yan, X. (2022, October). Customer preference identification from hotel online reviews: A neural network based fine-grained sentiment analysis. *Computers & Industrial Engineering*, *172*, 108648. doi: 10.1016/j.cie.2022.108648

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, July). Language (technology) is power: A critical survey of “bias” in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017, June). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. doi: 10.1162/tacl_a_00051

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. doi: 10.1016/j.jocs.2010.12.007

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of the 2019 World Wide Web Conference* (pp. 491–500). San Francisco USA: ACM. doi: 10.1145/3308560.3317593

Brown, S. L., & Eisenhardt, K. M. (1995). Product development: past research, present findings, and future directions. *Academy of Management Review*, *20*(2), 343–378. doi: 10.2307/258850

Buolamwini, J., & Gebru, T. (2018, February). Gender shades: intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). PMLR.

Burnap, A., Hauser, J. R., & Timoshenko, A. (2023). Product aesthetic design: A machine learning augmentation. *Marketing Science*, *42*(6), 1029–1056. doi: 10.1287/mksc.2022.1429

Cai, M., Yang, W., Du, Y., Tan, Y., & Lu, X. (2025, September). Automatic requirements elicitation from user-generated content: A review of data, methods, and representations. *Engineering Applications of Artificial Intelligence*, *156*, 111110. doi: 10.1016/j.engappai.2025.111110

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017, April). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. doi: 10.1126/science.aal4230

Chang, Y.-T., Yang, H.-R., & Chen, C.-M. (2022). Analysis on improving the application of machine learning in product development. *Journal of Supercomputing*, *78*(10), 12435–12460. doi: 10.1007/s11227-022-04344-3

Chatterjee, S., Ghatak, A., Nikte, R., Gupta, S., & Kumar, A. (2023, January). Measuring SERVQUAL dimensions and their importance for customer-satisfaction using online reviews: A text mining approach. *Journal of Enterprise Information Management*, *36*(1), 22–44. doi: 10.1108/JEIM-06-2021-0252

Cheligeer, C., Huang, J., Wu, G., Bhuiyan, N., Xu, Y., & Zeng, Y. (2022). Machine learning in requirements elicitation: A literature review. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *36*, e32. doi: 10.1017/S0890060422000166

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., … Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*(6), 719–742. doi: 10.1038/s41551-023-01056-8

Chew, O., Lin, H.-T., Chang, K.-W., & Huang, K.-H. (2024, March). Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In Y. Graham & M. Purver (Eds.), *Findings of the Association for Computational Linguistics: EACL 2024* (pp. 1013–1025). St. Julian's, Malta: Association for Computational Linguistics.

Chi, J., Shand, W., Yu, Y., Chang, K.-W., Zhao, H., & Tian, Y. (2022, December). Conditional supervised contrastive learning for fair text classification. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 2736–2756). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.199

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. doi: 10.1186/s12864-019-6413-7

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. doi: 10.1089/big.2016.0047

Coates, A., Ng, A., & Lee, H. (2011, April). An analysis of single-layer networks in unsupervised feature learning. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the Fourteenth*

International Conference on Artificial Intelligence and Statistics (Vol. 15, pp. 215–223). Fort Lauderdale, FL, USA: PMLR.

Cong, Y., Yu, S., Chu, J., Su, Z., Huang, Y., & Li, F. (2023, April). A small sample data-driven method: User needs elicitation from online reviews in new product iteration. *Advanced Engineering Informatics*, *56*, 101953. doi: 10.1016/j.aei.2023.101953

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *Arxiv Preprint Arxiv:1911.02116*.

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society*, *1*(1), 104–126. doi: 10.1007/BF03177550

Csanády, B., Muzsai, L., Vedres, P., Nádasdy, Z., & Lukács, A. (2024). LlamBERT: Large-scale low-cost data annotation in NLP. *Arxiv Preprint Arxiv:2403.15938*.

Cubric, M. (2020). Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study. *Technology in Society*, *62*, 101257. doi: 10.1016/j.techsoc .2020.101257

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120–128). Atlanta GA USA: ACM. doi: 10.1145/3287560.3287572

de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, *34*(14), 11751–11768. doi: 10.1007/s00521-022-07067-x

De Araújo, A. F., & Marcacini, R. M. (2021, March). RE-BERT: Automatic extraction of software requirements from app reviews using BERT language model. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (pp. 1321–1327). Virtual Event Republic of Korea: ACM. doi: 10.1145/3412841.3442006

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2021, August). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137. doi: 10.1609/icwsm.v7i1.14432

Delobelle, P., Tokpo, E., Calders, T., & Berendt, B. (2022, July). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1693–1706). Seattle, United States: Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.122

Denizci Guillet, B., & Kucukusta, D. (2016). Spa market segmentation according to customer preference. *International Journal of Contemporary Hospitality Management*, *28*(2), 418–434. doi: 10.1108/IJCHM-07-2014-0374

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Arxiv Preprint Arxiv:1810.04805*.

Dhamodharan, B. (2025, January). *AI agents: The next frontier in intelligent automation.* Forbes.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923. doi: 10.1162/089976698300017197

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on*

AI, Ethics, and Society (pp. 67–73). New Orleans LA USA: ACM. doi: 10.1145/3278721.3278729

Du, K., Xing, F., Mao, R., & Cambria, E. (2024, April). Financial sentiment analysis: techniques and applications. *ACM Computing Surveys*, *56*(9), 1–42. doi: 10.1145/3649451

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). Cambridge, Massachusetts and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2090236.2090255

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024, November). Cognitive bias in decision-making with LLMs. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 12640–12653). Miami, Florida, USA: Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.739

Ettrich, O., Stahlmann, S., Leopold, H., & Barrot, C. (2024). Automatically identifying customer needs in user-generated content using token classification. *Decision Support Systems*, *178*, 114107. doi: 10.1016/j.dss.2023.114107

Fan, W., & Davidson, I. (2007). On sample selection bias and its efficient correction via model averaging and unlabeled examples. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)* (pp. 320–331). doi: 10.1137/1.9781611972771.29

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, March). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37. doi: 10.1609/aimag.v17i3.1230

Feldman, R., & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 112–117). Montréal, Québec, Canada: AAAI Press.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*(1), 27–38. doi: 10.1016/j.patrec.2008.08.010

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329–338). Atlanta, GA, USA and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3287560.3287589

Gamzu, I., Gonen, H., Kutiel, G., Levy, R., & Agichtein, E. (2021). Identifying helpful sentences in product reviews. *Arxiv Preprint Arxiv:2104.09792*.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., & Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 219–226). Honolulu, HI, USA and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3306618.3317950

Garvie, C., Bedoya, A., & Frankle, J. (2016). *The perpetual line-up: Unregulated police face recognition in america* (Tech. Rep.). Georgetown Law Center on Privacy & Technology.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *Arxiv Preprint Arxiv:2303.15056*.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., & Lopez, A. (2021, August). Intrinsic bias metrics do not correlate with application bias. In C. Zong, F. Xia, W. Li, &

R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 1: Long Papers)* (pp. 1926–1940). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150

Gonzalez, M. E. (2019). Improving customer satisfaction of a healthcare facility: Reading the customers' needs. *Benchmarking: An International Journal*, *26*(3), 854–870. doi: 10.1108/BIJ-01-2017-0007

Griffin, A., & Hauser, J. R. (1993, February). The voice of the customer. *Marketing Science*, *12*(1), 1–27. doi: 10.1287/MKSC.12.1.1

Griffin, A., Price, R. L., Maloney, M. M., Vojak, B. A., & Sim, E. W. (2009). Voices from the field: How exceptional electronic industrial innovators innovate. *Journal of Product Innovation Management*, *26*(2), 222–240. doi: 10.1111/j.1540-5885.2009.00347.x

Guo, J., Tan, R., Sun, J., Ren, J., Wu, S., & Qiu, Y. (2016). A needs analysis approach to product innovation driven by design. *Procedia CIRP*, *39*, 39–44. doi: 10.1016/j.procir.2016.01.163

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. *Arxiv Preprint Arxiv:2004.10964*.

Guzman, E., Ibrahim, M., & Glinz, M. (2017, September). A little bird told me: Mining tweets for requirements and software evolution. In *2017 IEEE 25th International Requirements Engineering Conference (RE)* (pp. 11–20). Lisbon, Portugal: IEEE. doi: 10.1109/RE.2017.88

Haddaway, N. R., Grainger, M. J., & Gray, C. T. (2021). *Citationchaser: An R package and shiny app for forward and backward citations chasing in academic searching.* Retrieved 2025-08-25, from https://zenodo.org/record/4543513 doi: 10.5281/ZENODO.4543513

Hammersley, M., & Gomm, R. (1997). Bias in social research. *Sociological Research Online*, *2*(1), 7–19. doi: 10.5153/sro.55

Han, Y., Bruggeman, R., Peper, J., Chehade, E. C., Marion, T., Ciuccarelli, P., & Moghaddam, M. (2023). Extracting latent needs from online reviews through deep learning based language model. *Proceedings of the Design Society*, *3*, 1855–1864. doi: 10.1017/pds.2023.186

Han, Y., & Moghaddam, M. (2021). Eliciting attribute-level user needs from online reviews with deep language models and information extraction. *Journal of Mechanical Design*, *143*(6), 061403. doi: 10.1115/1.4048819

Han, Y., & Moghaddam, M. (2024, December). Design knowledge as attention emphasizer in large language model-based sentiment analysis. *Journal of Computing and Information Science in Engineering*, *25*(2), 21007. doi: 10.1115/1.4067212

Hanski, J., Reunanen, M., Kunttu, S., Karppi, E., Lintala, M., & Nieminen, H. (2014). Customer observation as a source of latent customer needs and radical new ideas for product-service systems. In J. Lee, J. Ni, J. Sarangapani, & J. Mathew (Eds.), *Engineering Asset Management 2011* (pp. 395–407). London: Springer London.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*.

Harrel, G. D. (1978). Book review: Modern marketing: Principles and practice. *Journal of Marketing*, *42*(4), 105–105. Retrieved 2025-08-26, from http://www.jstor.org/stable/1250103 doi: 10.1177/002224297804200426

Harzing, A.-W. (2016). *Publish or Perish.* Retrieved from https://harzing.com/

resources/publish-or-perish

Hasso, H., Fischer-Starcke, B., & Geppert, H. (2024). Quest-RE question generation and exploration strategy for requirements engineering. In *2024 IEEE 32nd International Requirements Engineering Conference Workshops (REW)* (pp. 1–9). Reykjavik, Iceland: IEEE. doi: 10.1109/REW61692.2024.00006

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *Arxiv Preprint Arxiv:2006.03654*.

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 507–517). Montréal Québec Canada: International World Wide Web Conferences Steering Committee. doi: 10.1145/2872427.2883037

He, Z., Ribeiro, M. T., & Khani, F. (2023, July). Targeted data generation: Finding and fixing model weaknesses. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers)* (pp. 8506–8520). Toronto, Canada: Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.474

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, *47*(1), 153–161. doi: 10.2307/1912352

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., … Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *Arxiv*, *abs/1712.409*.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1–12). Cognitive Science Society.

Hrinchenko, Y., Robul, Y., & Zalubinska, L. (2018, December). Development of price strategies to support brand positioning: Strategic issues for marketing policies. *Economic Innovations*, *20*(4(69)), 44–54. doi: 10.31520/ei.2018.20.4(69).44-54

Huang, S., Qin, H., Chan, T.-T., & Wang, Y. (2025, May). Identifying novel customer needs from user-generated content for product development using pre-trained language model. *Journal of Engineering Design*, 1–21. doi: 10.1080/09544828.2025.2504850

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. *Yale Journal of Law and Technology*, *18*, 148–216.

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., & Prabhakaran, V. (2022). Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1859–1876). Seoul Republic of Korea: ACM. doi: 10.1145/3531146.3533233

Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J., Flautner, K., … Mars, J. (2024). Scaling down to scale up: A cost-benefit analysis of replacing OpenAI's LLM with open source slms in production. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (pp. 280–291). Indianapolis, IN, USA: IEEE. doi: 10.1109/ISPASS61541.2024.00034

Islam, M. A., Kaium, M. A., Zahan, I., & Rahman, M. S. (2024, February). Does user-generated content trigger university graduates' online purchase intention? Mediating role of brand image. *Asian Management and Business Review*, 105–121. doi: 10.20885/AMBR.vol4.iss1.art7

Iswari, N. M. S., & Putra, I. G. J. E. (2023, June). Analysis of user-generated content in visitor reviews of tourist attractions using semantic similarity. *Ultimatics : Jurnal Teknik Informatika*, *15*(1), 59–64. doi: 10.31937/ti.v15i1.3139

Jhamtani, H., Chhaya, N., Karwa, S., Varshney, D., Kedia, D., & Gupta, V. (2015). Identifying suggestions for improvement of product features from online product reviews. In T.-Y. Liu, C. N. Scollon, & W. Zhu (Eds.), *Social Informatics* (Vol. 9471, pp. 112–119). Cham: Springer International Publishing. doi: 10.1007/978-3-319-27433-1_8

Kano, N., Seraku, N., Takahashi, F., & Tsuji, S.-i. (1984). Attractive quality and must-be quality. *Journal of the Japanese Society for Quality Control*, *14*(2), 147–156. doi: 10.20684/quality.14.2_147

Kärkkäinen, H., Piippo, P., Puumalainen, K., & Tuominen, M. (2001). Assessment of hidden and future customer needs in finnish business-to-business companies. *R&D Management*, *31*(4), 391–407. doi: 10.1111/1467-9310.00227

Kashi, M., Lahmiri, S., & Mohamed, O. A. (2025). Comprehensive analysis of transformer networks in identifying informative sentences containing customer needs. *Expert Systems with Applications*, *273*, 126785. Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417425004075 doi: https://doi.org/10.1016/j.eswa.2025.126785

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2019, August). Managing marketing decision-making with sentiment analysis: An evaluation of the main product features using text data mining. *Sustainability*, *11*(15), 4235. doi: 10.3390/su11154235

Kaur, K., & Kaur, P. (2023, October). MNoR-BERT: multi-label classification of non-functional requirements using BERT. *Neural Computing and Applications*, *35*(30), 22487–22509. doi: 10.1007/s00521-023-08833-1

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, July). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 2564–2572). PMLR.

Kilroy, D., Healy, G., & Caton, S. (2022). Using machine learning to improve lead times in the identification of emerging customer needs. *IEEE Access : Practical Innovations, Open Solutions*, *10*, 37774–37795.

Kilroy, D., Healy, G., & Caton, S. (2024, August). Prediction of future customer needs using machine learning across multiple product categories. *PLOS One*, *19*(8), e0307180. doi: 10.1371/journal.pone.0307180

King, T. (2020). Percent of your data will be unstructured in five years. *Retrieved February*, *16*, 80.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Tech. Rep. No. EBSE-2007-01). EBSE Technical Report, Keele University and University of Durham.

Kocon, J., Radom, J., Kaczmarz-Wawryk, E., Wabnic, K., Zajaczkowska, A., & Zasko-Zielinska, M. (2021, December). AspectEmo: multi-domain corpus of consumer reviews for aspect-based sentiment analysis. In *2021 International Conference on Data Mining Workshops (ICDMW)* (pp. 166–173). Auckland, New Zealand: IEEE. doi: 10.1109/ICDMW53433.2021.00027

Kovacs, M., Buryakov, D., & Kryssanov, V. (2021, September). An unsupervised approach for customer need assessment in E-commerce: A case study of japanese customer reviews. In *2021 6th International Conference on Cloud Computing and Internet of Things* (pp. 41–48). Okinawa Japan: ACM. doi: 10.1145/3493287.3493294

Kuehl, N., Mühlthaler, M., & Goutier, M. (2020). Supporting customer-oriented marketing with artificial intelligence: Automatically quantifying customer needs from social media. *Electronic*

Markets, *30*(2), 351–367. doi: 10.1007/s12525-019-00351-0

Kuehl, N., Scheurenbrand, J., & Satzger, G. (2016). NEEDMINING: Identifying micro blog data containing customer needs. In *Proceedings of the European Conference on Information Systems (ECIS)* (Vol. 185). Association for Information Systems (AIS).

Kühl, N., Scheurenbrand, J., & Satzger, G. (2020). Needmining: Identifying micro blog data containing customer needs. arXiv. doi: 10.48550/arXiv.2003.05917

Kumar, D., & Suthar, N. (2024). Ethical and legal challenges of AI in marketing: An exploration of solutions. *Journal of Information, Communication and Ethics in Society*, *22*(1), 124–144. doi: 10.1108/JICES-05-2023-0068

Kumar, D. T. S. (2020). Data mining based marketing decision support system using hybrid machine learning algorithm. *Journal of Artificial Intelligence and Capsule Networks*, *2*(3), 185–193. doi: 10.36548//jaicn.2020.3.006

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4069–4079). Long Beach, California, USA and Red Hook, NY, USA: Curran Associates Inc.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *Arxiv Preprint Arxiv:1909.11942*.

Law, K., Majava, J., Nuottila, J., & Haapasalo, H. (2014, January). Customer needs in market-driven product development: Product management and R&D standpoints. *Technology and Investment*, *5*(1), 16–25. doi: 10.4236/ti.2014.51003

Lee, J., Jeong, B., Yoon, J., & Song, C. H. (2023). Context-aware customer needs identification by linguistic pattern mining based on online product reviews. *IEEE Access : Practical Innovations, Open Solutions*, *11*, 71859–71872. doi: 10.1109/ACCESS.2023.3295452

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (oxford, England)*, *36*(4), 1234–1240.

Leonard, D., Rayport, J. F., & Others. (1997). Spark innovation through empathic design. *Harvard Business Review*, *75*, 102–115.

Li, C., Huang, L., Ge, J., Luo, B., & Ng, V. (2018, April). Automatically classifying user requests in crowdsourcing requirements engineering. *Journal of Systems and Software*, *138*, 108–123. doi: 10.1016/j.jss.2017.12.028

Li, H. (2017, September). Deep learning for natural language processing: Advantages and challenges. *National Science Review*, *5*(1), 24–26. doi: 10.1093/nsr/nwx110

Li, J., & Cao, B. (2022, July). Study on tourism consumer behavior and countermeasures based on big data. *Computational Intelligence and Neuroscience*, *2022*, 1–12. doi: 10.1155/2022/6120511

Li, M., Shi, L., Yang, Y., & Wang, Q. (2020, December). A deep multitask learning approach for requirements discovery and annotation from open forum. In *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering* (pp. 336–348). Virtual Event Australia: ACM. doi: 10.1145/3324884.3416627

Li, Q., Yang, Y., Li, C., & Zhao, G. (2023, December). Energy vehicle user demand mining method based on fusion of online reviews and complaint information. *Energy Reports*, *9*, 3120–3130. doi: 10.1016/j.egyr.2023.02.004

Liu, H., Jin, W., Karimi, H., Liu, Z., & Tang, J. (2021). The authors matter: Understanding and mitigating implicit bias in deep text classification. *Arxiv Preprint Arxiv:2105.02778*.

Liu, N. F., Kumar, A., Liang, P., & Jia, R. (2022). Are sample-efficient NLP models more robust?

Arxiv Preprint Arxiv:2210.06456.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *Arxiv Preprint Arxiv:1907.11692*.

Liu, Y., Zhang, H., Li, Z., & Miao, Y. (2024). Optimizing the utilization of large language models via schedule optimization: An exploratory study. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 84–95). Barcelona, Spain and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3674805.3686671

Luca, M. (2015). Chapter 12 - user-generated content and social media. In S. P. Anderson, J. Waldfogel, & D. Strömberg (Eds.), *Handbook of Media Economics* (Vol. 1, pp. 563–592). North-Holland. doi: 10.1016/B978-0-444-63685-0.00012-7

Ma, X., Xu, P., Wang, Z., Nallapati, R., & Xiang, B. (2019). Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-resource NLP (deeplo 2019)* (pp. 76–83). Hong Kong, China: Association for Computational Linguistics. doi: 10.18653/v1/D19-6109

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Portland, Oregon, USA: Association for Computational Linguistics.

Mahdi, H. F., Gupta, L. K., Choudhury, T., & Bansal, N. (2022, October). Idea mining from online reviews using transformation-based natural language processing tasks. In *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 894–899). Ankara, Turkey: IEEE. doi: 10.1109/ISMSIT56059.2022.9932785

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Matzler, K., & Hinterhuber, H. H. (1998). How to make product development projects more successful by integrating kano's model of customer satisfaction into quality function deployment. *Technovation*, *18*(1), 25–38. doi: 10.1016/s0166-4972(97)00072-2

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular pdp networks and distributed lexicon. *Cognitive Science*, *15*(3), 343–399. doi: 10.1207/s15516709cog1503_2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Arxiv Preprint Arxiv:1301.3781*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40. doi: 10.1145/3439726

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. doi: 10.1016/j.patcog.2011.06.019

Mujtaba, D. F., & Mahapatra, N. R. (2019). Ethical considerations in AI-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)* (pp. 1–7). Medford, MA, USA: IEEE. doi: 10.1109/ISTAS48451.2019.8937920

Nadeau, C., & Bengio, Y. (1999). Inference for the generalization error. *Advances in Neural Information Processing Systems*, *12*.

Naeem, M., & Ozuem, W. (2022a, December). Understanding misinformation and rumors that generated panic buying as a social practice during COVID-19 pandemic: Evidence from twitter,

YouTube and focus group interviews. *Information Technology & People*, *35*(7), 2140–2166. doi: 10.1108/ITP-01-2021-0061

Naeem, M., & Ozuem, W. (2022b, December). Understanding misinformation and rumors that generated panic buying as a social practice during COVID-19 pandemic: Evidence from twitter, YouTube and focus group interviews. *Information Technology & People*, *35*(7), 2140–2166. doi: 10.1108/ITP-01-2021-0061

Narver, J. C., Slater, S. F., & MacLachlan, D. L. (2004). Responsive and proactive market orientation and new-product success. *Journal of Product Innovation Management*, *21*(5), 334–347. doi: 10.1111/j.0737-6782.2004.00086.x

Nozza, D., Volpetti, C., & Fersini, E. (2019). Unintended bias in misogyny detection. In *Ieee/wic/acm International Conference on Web Intelligence* (pp. 149–155). Thessaloniki Greece: ACM. doi: 10.1145/3350546.3352512

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. doi: 10.1126/science.aax2342

Ochoa, X., & Duval, E. (2008). Quantitative analysis of user-generated content on the web. In *Proceedings of Webevolve2008: Web Science Workshop at WWW2008* (pp. 1–8).

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on twitter. *Internet Interventions*, *2*(2), 183–188. doi: 10.1016/j.invent.2015.03.005

O'Hern, M. S., & Kahle, L. R. (2013). The empowered customer: User-generated content and the future of marketing. *Global Economics and Management Review*, *18*(1), 22–30. doi: 10.1016/s2340-1540(13)70004-5

Olteanu, A., Talamadupula, K., & Varshney, K. R. (2017). The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 405–406). Troy, New York, USA and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3091478.3098871

OpenAI. (2023). *OpenAI models overview.*

OpenAI. (2023b). *OpenAI API reference.* https://platform. openai.com/docs/api-reference/chat/create.

Otto, K., & Wood, K. (2001). *Product design: Techniques in reverse engineering and new product development: Techniques in reverse engineering and new product development*. United States: Prentice Hall.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (clinical Research Ed.)*, *372*, n71. doi: 10.1136/bmj.n71

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL* (pp. 115–124). Ann Arbor, Michigan: Association for Computational Linguistics. doi: 10.3115/1219840.1219855

Peng, B., Chersoni, E., Hsu, Y.-Y., & Huang, C.-R. (2021). Is domain adaptation worth your investment? Comparing BERT and FinBERT on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing* (pp. 37–44). Punta Cana, Dominican Republic: Association for Computational Linguistics. doi: 10.18653/v1/2021.econlp-1.5

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014*

Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. doi: 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Corr*, *abs/1802.5365*.

Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, *32*(23), 17309–17320. doi: 10.1007/s00521-020-05102-3

Prabhakaran, V., Hutchinson, B., & Mitchell, M. (2019, November). Perturbation sensitivity analysis to detect unintended model biases. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5740–5745). Hong Kong, China: Association for Computational Linguistics. doi: 10.18653/v1/D19-1578

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google translate. *Neural Computing and Applications*, *32*(10), 6363–6381. doi: 10.1007/s00521-019-04144-6

Proserpio, D., Hauser, J. R., Liu, X., Amano, T., Burnap, A., Guo, T., ... Others (2020). Soul and machine (learning). *Marketing Letters*, *31*(4), 393–404. doi: 10.1007/s11002-020-09538-4

Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.. Retrieved from https://api.semanticscholar.org/CorpusID:49313245

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). Barcelona, Spain and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3351095.3372828

Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 873–880). Montreal, Quebec, Canada and New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1553374.1553486

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Arxiv Preprint Arxiv:1908.10084*.

Reiss, M. V. (2023). Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. *Arxiv Preprint Arxiv:2304.11085*.

Reshmi, s., & Balakrishnan, K. (2018, February). Empowering chatbots with business intelligence by big data integration. *International Journal of Advanced Research in Computer Science*, *9*(1), 627–631. doi: 10.26483/ijarcs.v9i1.5398

Reuther, K. K. (2022). *Shrink it and pink it: Gender bias in product design.* https://www.sir.advancedleadership.harvard.edu/articles/shrink-it-and-pink-it-gender-bias-product-design.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. doi: 10.1162/tacl_a_00349

Saldaña, J. (2021). The coding manual for qualitative researchers. *The Coding Manual for Qualitative Researchers*, 1–440.

Salminen, J., Jung, S.-G., & Jansen, B. J. (2021, October). Manual and automatic methods for user needs detection in requirements engineering: Key concepts and challenges. In *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1–7). Mauritius, Mauritius: IEEE. doi: 10.1109/ICECCME52200.2021.9591046

Salminen, J., Mustak, M., Corporan, J., Jung, S.-g., & Jansen, B. J. (2022). Detecting pain points from user-generated social media posts using machine learning. *Journal of Interactive Marketing*, *57*(3), 517–539. doi: 10.1177/10949968221095556

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Arxiv Preprint Arxiv:1910.01108*.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1163

Scholarcy. (2021). Retrieved from https://www.scholarcy.com

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019, April). Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, *7*(2), e12239. doi: 10.2196/12239

Smets, L. P. M., Langerak, F., & Rijsdijk, S. A. (2013). Shouldn't customers control customized product development? *Journal of Product Innovation Management*, *30*(6), 1242–1253. doi: 10.1111/jpim.12057

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Seattle, Washington, USA: Association for Computational Linguistics. doi: 10.18653/v1/D13-1170

Stahlmann, S., Ettrich, O., Kurka, M., & Schoder, D. (2023). What do customers say about my products? Benchmarking machine learning models for need identification. In *Proc. of the HICSS*. doi: 10.24251/HICSS.2023.264

Stahlmann, S., Ettrich, O., & Schoder, D. (2022). Deep learning enabled consumer research for product development. In *ECIS 2022 Research-in-progress Papers*. AIS Electronic Library (AISeL).

Steed, R., Panda, S., Kobren, A., & Wick, M. (2022, May). Upstream mitigation is *not* all you need: Testing the bias transfer hypothesis in pre-trained language models. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers)* (pp. 3524–3542). Dublin, Ireland: Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.247

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18* (pp. 194–206). Springer.

Thornton, C. (2002). *Truth from trash: How learning makes sense*. Mit Press.

Timoshenko, A., & Hauser, J. R. (2019, January). Identifying customer needs from user-generated content. *Marketing Science*, *38*(1), 1–20. doi: 10.1287/mksc.2018.1123

Timoshenko, A., Mao, C., & Hauser, J. R. (2025). *Can large language models extract customer*

needs as well as professional analysts? arXiv. doi: 10.48550/ARXIV.2503.01870

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Venkit, P. N., Srinath, M., & Wilson, S. (2022, October). A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 1324–1332). Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

von Hippel, E. (1986). Lead users: A source of novel product concepts. *Management Science*, *32*(7), 791–805. doi: 10.1287/mnsc.32.7.791

Wang, C., Jiang, X., Li, Q., Hu, Z., & Lin, J. (2024). *Leveraging llm agents to extract customer needs from user-generated content.* SSRN. doi: 10.2139/ssrn.4973155

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., ... Yu, P. S. (2023). Generalizing to unseen domains: a survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 8052–8072. doi: 10.1109/TKDE.2022.3178128

Wang, T., Sridhar, R., Yang, D., & Wang, X. (2021). Identifying and mitigating spurious correlations for improving robustness in nlp models. *Arxiv Preprint Arxiv:2110.07736*.

Wang, Y., Mo, D. Y., & Tseng, M. M. (2018). Mapping customer needs to design parameters in the front end of product design by applying deep learning. *CIRP Annals*, *67*(1), 145–148. doi: 10.1016/j.cirp.2018.04.018

Wei, D., Ramamurthy, K. N., & Calmon, F. (2020, August). Optimized score transformation for fair classification. In S. Chiappa & R. Calandra (Eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (Vol. 108, pp. 1673–1683). PMLR.

Wei, W., Hao, C., & Wang, Z. (2025, May). User needs insights from UGC based on large language model. *Advanced Engineering Informatics*, *65*, 103268. doi: 10.1016/j.aei.2025.103268

Wiegers, K. E., & Beatty, J. (2013). *Software requirements*. Pearson Education.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE)* (pp. 38:1–38:10). New York, NY, USA: ACM. doi: 10.1145/2601248.2601268

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Online: Association for Computational Linguistics.

Wu, H. (2023, September). Leveraging data analytics and consumer insights for targeted marketing campaigns and personalized customer experiences. *Journal of World Economy*, *2*(3), 33–44. doi: 10.56397/JWE.2023.09.05

Xiao, Y., Li, C., Thürer, M., Liu, Y., & Qu, T. (2022, September). User preference mining based on fine-grained sentiment analysis. *Journal of Retailing and Consumer Services*, *68*, 103013. doi: 10.1016/j.jretconser.2022.103013

Xie, Y., Yeoh, W., & Wang, J. (2024). How self-selection bias in online reviews affects buyer satisfaction: A product type perspective. *Decision Support Systems*, *181*, 114199. doi: 10 .1016/j.dss.2024.114199

Yan, X., Li, Y., & Fan, W. (2017, November). Identifying domain relevant user generated content through noise reduction: A test in a chinese stock discussion forum. *Information Discovery and Delivery*, *45*(4), 181–193. doi: 10.1108/IDD-04-2017-0043

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yin, C., Jiang, C., Jain, H. K., Liu, Y., & Chen, B. (2023, September). Capturing product/service improvement ideas from social media based on lead user theory. *Journal of Product Innovation Management*, *40*(5), 630–656. doi: 10.1111/jpim.12676

Young, R. R. (2004). *The requirements engineering handbook*. Artech House.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning* (p. 114). Banff, Alberta, Canada: ACM Press. doi: 10.1145/1015330.1015425

Zhan, Y., Tan, K. H., Li, Y., & Tse, Y. K. (2018, November). Unlocking the power of big data in new product development. *Annals of Operations Research*, *270*(1-2), 577–595. doi: 10.1007/s10479-016-2379-x

Zhang, G., Bai, B., Zhang, J., Bai, K., Zhu, C., & Zhao, T. (2020, July). Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4134–4145). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.380

Zhang, K., Lin, K.-Y., Wang, J., Ma, Y., Li, H., Zhang, L., ... Feng, L. (2023, August). UNISON framework for user requirement elicitation and classification of smart product-service system. *Advanced Engineering Informatics*, *57*, 101996. doi: 10.1016/j.aei.2023.101996

Zhang, M., Fan, B., Zhang, N., Wang, W., & Fan, W. (2021). Mining product innovation ideas from online reviews. *Information Processing & Management*, *58*(1), 102389. doi: 10.1016/j.ipm.2020.102389

Zhang, M., Sun, L., Li, Y., Wang, G. A., & He, Z. (2023). Using supplementary reviews to improve customer requirement identification and product design development. *Journal of Management Science and Engineering*, *8*(4), 584–597. doi: 10.1016/j.jmse.2023.03.001

Zhang, Z., Dou, Y., Xu, X., & Tan, Y. (2024, April). A PTM-based framework for enhanced user requirement classification in product design. *Electronics*, *13*(8), 1458. doi: 10.3390/electronics13081458

Zhao, Q., Zhao, W., Guo, X., Zhang, K., & Yu, M. (2022, December). A dynamic customer requirement mining method for continuous product improvement. *Autonomous Intelligent Systems*, *2*(1), 14. doi: 10.1007/s43684-022-00032-4

Zhao, Y., & Tang, Q. (2021, April). Analysis of influencing factors of social mental health based on big data. *Mobile Information Systems*, *2021*, 1–8. doi: 10.1155/2021/9969399

Zhou, F., Ayoub, J., Xu, Q., & Jessie Yang, X. (2020). A machine learning approach to customer needs analysis for product ecosystems. *Journal of Mechanical Design*, *142*(1), 11101. doi: 10.1115/1.4044435

Zhou, F., Jiao, R. J., & Linsey, J. S. (2015). Latent customer needs elicitation by use case analogical reasoning from sentiment analysis of online product reviews. *Journal of Mechanical Design*, *137*(7), 71401. doi: 10.1115/1.4030159

Zmigrod, R., Mielke, S. J., Wallach, H., & Cotterell, R. (2019, July). Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1651–1661). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1161

Zotero. (2007). Retrieved from www.zotero.org