Integration of Inconsistency and Content Interaction with Deep Learning to Detect Fake Reviews

Kiana Sharifpour

A Thesis

in

The Department

of

Supply Chain & Business Technology Management

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Business Analytics and Technology Management at

Concordia University

Montréal, Québec, Canada

December 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared,

By: Kiana Sharifpour

Entitled: Integration of Inconsistency and Content Interaction with Deep Learning to Detect Fake Reviews

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

Master of Business Analytics and Technology Management

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

Signed by the final Examining Committee:

Dr. Arman Sadreddin	Chair
Dr. Danielle Morin	Examiner
Dr. Mahdi Mirhoseini	Examiner
Dr. Salim Lahmiri	Supervisor

Approved by: Dr. Suchit Ahuja, Graduate Program Director

Date: December 06, 2024

Dean: Dr. Anne-Marie Croteau

ABSTRACT

Integration of Inconsistency and Content Interaction with Deep Learning to Detect Fake Reviews

Kiana Sharifpour

In this study, the challenge of detecting fake reviews in e-commerce is addressed through the application of natural language processing and deep learning techniques. The paper introduces two frameworks designed to identify fraudulent reviews, a critical concern due to their impact on consumer behavior and market dynamics. Central to this study is the exploitation of rating-sentiment inconsistency (RSI), a nuanced textual feature indicative of potential deception, aimed at enhancing the detection of fake content. Using a dataset of Amazon reviews, the paper evaluates two distinct approaches. The first method integrates RSI with word embeddings, specifically GloVe and Word2Vec, yielding an accuracy improvement of 3.07% for GloVe and 0.67% for Word2Vec, demonstrating the effectiveness of BiGRUs in capturing the sequential nature of textual data. The second method incorporates inconsistency features into Doc2Vec representations, achieving a 1.55% increase in accuracy compared to models without this feature. Both methodologies benefit from grid search optimization to fine-tune hyperparameters, enhancing model performance significantly. These combination methods not only underscore the importance of content-based feature integration but also demonstrate the practical application of inconsistency metrics in fake review detection. The observed improvements in model accuracy confirm the effectiveness of the proposed frameworks, providing new insights into enhancing online review system integrity and advancing natural language processing in commercial settings.

Keywords: Fake review detection, review inconsistency, deep learning (DL), word embeddings (WE), natural language processing (NLP), text classification

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to Dr. Salim Lahmiri, my thesis supervisor, whose invaluable guidance, and steadfast support have been instrumental throughout this transformative journey. His unwavering dedication and profound expertise have profoundly influenced the development of this work.

I would also like to express my gratitude to all those who have supported me throughout this journey in various capacities. I am profoundly thankful.

Table of Contents

List of figures

List of Tables

Acronyms

UGC	User-Generated Content
WOM	Word-of-Mouth
AI	Artificial Intelligence
NLP	Natural Language Processing
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
GRU	Gated Recurrent Unit
BiGRU	Bidirectional Gated Recurrent Unit
CNN	Convolutional Neural Network
RSI	Rating-Sentiment Inconsistency
SVM	Support Vector Machine
ML	Machine Learning
DL	Deep Learning
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative Pre-trained Transformer

Chapter 1 INTRODUCTION

The rise of e-commerce technologies has profoundly altered consumer purchasing behavior, with online reviews and user-generated content (UGC) playing a central role. Online reviews are viewed as complimentary aid for sales and hold a significant influence on consumers' buying choices (S. Wang et al., 2023). They are vital in e-commerce, offering shoppers credible insights and social proof about product quality, thereby reducing uncertainty in online purchases (Y. Wang, Zamudio, et al., 2023). Online sales represent a growing share of the retail market, driven in part by word-of-mouth (WOM) that amplifies consumer awareness and product sales. Consumers depend on WOM, particularly when assessing products that are challenging to evaluate before purchase. The advent of the internet has expanded WOM from local to global networks. Consumers increasingly turn to product scores and reviews to make informed choices, relying on the wisdom of the crowd (Harrison-Walker & Jiang, 2023).

However, the reliability of online reviews has come under scrutiny due to the prevalence of fake reviews. Fake reviews are deceptive reviews written to artificially inflate or deflate the ratings of products and services. These reviews can mislead consumers, damage business reputations, and skew market dynamics. The detection of fake reviews has then become a crucial aspect of maintaining the integrity of e-commerce platforms. A variety of machine learning techniques have been employed to detect fake reviews, including supervised learning, clustering, and graph-based methods. In the supervised learning approach, reviews are labelled as fake or real, and various features, such as the presence of minimal information about the reviewer, short reviews, similarity in review content, focus on personal information, timeframe of posting reviews, and excessive use of positive and negative words, are used to distinguish between the two. Meanwhile, in the unsupervised method, unlabeled reviews are grouped into clusters based on their similarity (Nagi Alsubari et al., 2022).

By using a variety of textual and behavioral features, natural language processing (NLP) models can effectively detect fake reviews and provide more reliable information for consumers (Birim et al., 2022). However, due to privacy concerns and dataset availability, access to behavioral features of customers might not always be feasible nor ethical. Therefore, achieving high performance in fake review detection by merely relying on content and textual features of reviews would be a valuable achievement and important milestone.

1.1 Real-world Impact and Economic Implications

Fake reviews significantly impact both consumers and businesses, often leading to misguided purchasing decisions and reputational damage. Consumers can be misled into buying inferior products based on deceptive positive reviews, resulting in dissatisfaction and a potential loss of trust in online reviews. For instance, the skincare brand Sunday Riley faced significant backlash when it was revealed that employees were instructed to write fake positive reviews on Sephora. This not only damaged the brand's reputation but also led to legal and financial repercussions (He et al., 2022).

Another example is the accessory manufacturer Aukey, which was banned from Amazon for paying customers to leave positive reviews. Such practices highlight the risk businesses take when they attempt to manipulate their reputation through fake reviews (He et al., 2022). This loss of consumer trust is critical, as a study found that 81% of consumers would avoid a brand if they suspected it of using fake reviews, and 48% would leave a negative review (Schuman, 2022; *Study Examines the Impact of Fake Online Reviews on Sales*, 2022).

The economic impact of fake reviews is substantial. Based on 2021 statistics, fake reviews were estimated to influence $152 billion in online spending each year. In major e-commerce markets, the impact was also significant: $791 billion in the US, $6.4 billion in Japan, $5 billion in the UK, $2.3 billion in Canada, and $900 million in Australia (Marciano, 2021).

Fake reviews distort market dynamics by providing unfair competitive advantages to businesses that use them and disadvantaging their competitors. This manipulation leads to a misallocation of resources and a distorted competitive landscape. A study by Alma Economics estimated that fake reviews cause between £50 million to £312 million in total harm to UK consumers, highlighting the substantial economic detriment they cause (“Fake Online Reviews Research,” 2023).

1.2 Generative AI and Fake Content

The advent of generative AI has significantly impacted the proliferation of fake content. Advanced models like GPT-3 and GPT-4 can create highly realistic text that mimics human writing styles, making

it challenging for traditional verification methods to distinguish between authentic and fabricated information. This technology can generate fake news, reviews, and social media posts at scale, potentially exacerbating misinformation and deception. The ability of generative AI to produce convincing deepfakes in text, audio, and video formats further complicates efforts to verify content authenticity. While these models hold promise for positive applications, their misuse underscores the need for robust AI ethics, detection technologies, and regulatory frameworks to mitigate the spread of fake content (Floridi & Chiriatti, 2020; Solaiman et al., n.d.). To counter the rise of AI-generated fake content, researchers are developing AI-driven detection tools that analyze linguistic patterns, metadata, and other features to identify fabricated information. Adversarial training, which involves exposing detection models to both real and fake data, enhances the robustness of these systems (Zellers et al., 2020).

Several studies have focused on the detection of AI-generated scientific abstracts, highlighting the effectiveness of various machine learning models in this context. For instance, one study utilized an ensemble of transformer models to detect generated scientific papers, demonstrating the potential of advanced AI techniques in identifying fabricated academic content (Weber-Wulff et al., 2023). Similarly, research on AI-generated fake reviews has explored the application of neural networks and other machine learning models. These studies emphasize the importance of leveraging both textual and behavioral features to improve detection accuracy. Despite promising results, these efforts are nascent, and further research is needed to refine these techniques and adapt them to the ever-evolving landscape of AI-generated content (Glazkova & Glazkov, 2022).

The ethical deployment of generative AI necessitates guidelines to prevent misuse. Policymakers are working on regulations to ensure accountability and transparency in AI applications, balancing innovation benefits with the need to curb malicious activities (Jobin et al., 2019).

1.3 Fake Reviews Legal Framework in Canada

In Canada, the Competition Act governs the use of fake reviews, prohibiting false or misleading representations to the public. Violations of this act can result in both criminal and civil penalties, including fines and imprisonment. For criminal offenses, individuals can face fines up to $200,000 and up to one year in prison for summary conviction, with higher fines and up to 14 years in prison for

indictable offenses. Civil penalties can reach up to $750,000 for individuals and $10 million for corporations on the first offense, with increased penalties for subsequent offenses. Businesses are required to actively monitor and manage their online reviews to ensure compliance. This includes removing or reporting fake reviews, as failure to do so, even if the business did not create the reviews, can result in significant penalties. The Competition Bureau actively enforces these regulations and encourages businesses to solicit genuine reviews ethically and educate employees on the legal implications of fake reviews (*Five-Star Fake Out*, 2022).

One prominent case involved Bell Canada, which was fined $1.25 million in 2015 after its employees were found to have posted positive reviews of Bell's apps without disclosing their employment. This case underscores the serious legal and reputational risks associated with fake reviews. Another example is the crackdown on HCSC Inc., which operated under the name "Mr. Hobbs Coffee," and was found to have paid for fake reviews to boost its online presence (*The Deceptive Marketing Practices Digest — Volume 1*, 2015).

Addressing the prevalence of fake reviews is crucial for maintaining the integrity and trustworthiness of e-commerce platforms. The legal framework in Canada, enforced by the Competition Bureau, aims to deter such deceptive practices through stringent penalties and active monitoring. By ensuring compliance and promoting genuine reviews, businesses can help protect consumers and maintain fair market competition source (*Fraud and Scams*, 2022).

1.4 Focus on Review Inconsistency

In our study, review inconsistency would be used as a complementary feature to detect fake reviews using textual features and deep learning classifiers. Inconsistency happens when different parts of a review or different reviews are in conflict with respect to specific dimensions (Shan et al., 2021). As identified by Harrison-Walker and Jiang (Harrison-Walker & Jiang, 2023), review composition nonconformity which focuses on the readability and grammatical correctness of reviews is a cue that a review is fake. Our focus on "rating-sentiment inconsistency" aligns with the review composition non-conformity aspect. Poor grammar and spelling errors are cues often used by consumers to identify fake reviews. In our case, the inconsistency arises from disparities between the review's content and its accompanying rating, reflecting a form of non-conformity in review composition. By addressing

inconsistencies in the composition of reviews, particularly in the context of RSI, our methodology is expected to enhance the detection of fake reviews. Table 1. shows four examples of fake reviews with high inconsistency between their sentiment and rating. As seen, there is no consistency between the written reviews and their corresponding ratings provided by the reviewers.

Table 1.1 Examples of Fake Reviews with High Rating-Sentiment Inconsistency in the Dataset

Rating	Review
1	I've had a previous version of this and it plays so much better. the only thing this has over the u818a was that the camera is better, but not 30 dollars better.
1	good product. i use this product but not very good. i am not satisfied hi I am rizvi2xxx(skypee name) contact me I am a professional amazon worker.
5	This doomed crystal skill shot glass is bad to the bone. Bad addition to my collection. Looks cool with colored beverages.
5	The tabletop is fine but the 300 chips included are made of very very light weight plastic and are worthless, so don't let the chips influence your purchase.

1.5 Theoretecial Frameworks

We adapt cognitive dissonance theory and truth-default theory to explain the relationship between review inconsistency and the prevalence of fake reviews (Abdulqader et al., 2022; Harmon-Jones & Mills, 2019). Cognitive dissonance theory suggests that individuals strive for consistency in their behaviors. Users who provide inconsistent information in their reviews and ratings experience cognitive dissonance, which leads to a state of psychological discomfort. To resolve this issue, users engage in various strategies, such as avoiding information and truth. This avoidance of truth and information-seeking may be a mechanism used by individuals to alleviate their cognitive dissonance. However, this strategy often leads to the creation of written reviews that are more likely to be fraudulent and inauthentic (Harmon-Jones & Mills, 2019). On the other hand, based on truth-default theory, reviews and ratings from a truthful person do not conflict with each other. A lack of consistency and

coherency in content dimensions is suspicious and may be a sign of deception. The conflict between the message said and the behavior of its writer can be considered as a deception cue indicating inauthenticity of the message (Abdulqader et al., 2022).

1.6 Contribution

Various review inconsistency metrics have been proposed in the literature among which RSI, derived from review rating and review content, and its impact on model performance will be discussed. Review rating has a major role on consumers' willingness to purchase a product. The existence of inconsistency in review ratings for a specific product, for instance, can be perceived as a concern for consumers towards the product's quality and affect product sales negatively. Therefore, addressing the prevalence of inconsistency among real and fake reviews and eliminating its source can play a major role for both sellers and platform profitability (Eslami & Ghasemaghaei, 2018).

Our primary contribution would be the integration of review inconsistency, specifically RSI, into fake review detection models using deep learning techniques and pretrained word embeddings. The following questions will be answered in this study:

1. How does the combination of RSI with GloVe and Word2Vec review representations affect the accuracy of fake review detection?
2. How does the integration of RSI values as new vocabulary items and combined with document embeddings influence the performance of fake review detection models?
3. How does the incorporation of RSI in the proposed frameworks enhance their effectiveness compared to traditional fake review detection models that do not use this feature?

To answer these questions, two combination methods for incorporating RSI with review representations learned from pre-trained word embeddings and document embeddings are proposed in the methodology section. The proposed methodology would prove beneficial in contexts where access to user behavior is constrained. In such scenarios, relying solely on textual context and the limited information about each dependent feature becomes pivotal for effectively classifying spam. The rating star feature is accessible in most large online review datasets, and extracting sentiment from review

content is feasible through various NLP-based methods. Therefore, RSI can potentially serve as easily obtainable and extractable features in numerous contexts. In summary, the main contributions of this study are as follows:

- Leveraging pre-trained word embeddings and document embeddings to detect fake reviews more effectively by introducing two novel combination methods for integrating inconsistency with review representations.
- Demonstration of how the integration of RSI with review representations enhances the accuracy of fake review detection models, especially in contexts where behavioral data is limited or unavailable.
- Highlighting the importance of relying on textual context and the innovative use of RSI as a complementary feature. This approach is crucial for scenarios where only the content of reviews and their ratings are accessible for analysis.

The contributions of this thesis are twofold. First, it addresses the limitation of previous works by integrating RSI, a novel feature that captures the disparity between the textual content of reviews and their corresponding star ratings. This metric adds a new dimension to fake review detection, complementing traditional textual and behavioral features.

Second, the thesis introduces two frameworks combining RSI with pre-trained word embeddings (such as GloVe and Word2Vec) and document embeddings (Doc2Vec), which offer significant improvements in classification accuracy. By leveraging BiGRU networks, our models are better equipped to capture sequential relationships in the review text, leading to more robust detection of fake reviews.

Compared to baseline models, the integration of RSI with advanced embedding techniques has shown substantial performance improvements, with up to 3.07% accuracy increase when using GloVe embeddings. This highlights the importance of including inconsistency metrics in the broader discussion of fake review detection and provides a strong foundation for future work in this evolving field.

1.7 Structure of Thesis

This dissertation is organized into seven comprehensive chapters, each building on the previous to provide a holistic view of the research on detecting fake reviews using natural language processing and deep learning methodologies.

Chapter 2 will provide an extensive review of existing research in the field of fake review detection. It will cover various natural language processing (NLP) and machine learning techniques previously utilized to identify fake reviews, highlighting the strengths and limitations of each approach.
In chapter 3, the focus will shift to the specific methodologies employed in this study to detect fake reviews on Amazon dataset. This chapter will detail the development and implementation of the proposed frameworks, including the integration of rating-sentiment inconsistency (RSI) with pre-trained word embeddings and document embeddings. Chapter 4 will describe the dataset used in the study. It will explain the rationale behind selecting the Amazon review dataset, detailing its characteristics and why it is suitable for the objectives of this research.
The results of the study will be presented in chapter 5. It will include an analysis of the performance of the proposed models, comparing them with baseline models to demonstrate the effectiveness of the new approaches.
The final chapter will provide a summary of the findings, discussing their implications for the field of fake review detection. It will also suggest directions for future research, identifying potential areas for further exploration to advance the integrity of online review systems.

Chapter 2 LITERATURE REVIEW

2.1 Inconsistency between rating and content

Extensive studies in various fields, such as recommender systems and review usefulness, have examined the disparity between review content and the assigned ratings. These investigations have delved into causes, precursors, outcomes, and the distribution of such discrepancies. Findings suggest that users exhibit greater consistency at extreme rating levels (Kotkov et al., 2022), although our research indicates higher inconsistency at these points as shown in Fig. 1. This contradiction calls for more research on the distribution of inconsistencies and their influence on various constructs.

Difficulty in assigning ratings is a cited reason for inconsistency (Kotkov et al., 2022). Furthermore, a high degree of inconsistency has been observed to compromise the credibility and helpfulness of online reviews (Y. Wang, Ngai, et al., 2023). In contrast to the majority of information pertaining to online reviews, which is frequently either inaccessible or necessitates explicit extraction, the inconsistency between ratings and textual sentiments—a construct derived from numerical ratings and textual sentiments—is consistently available and readily computable across online reviews. This characteristic makes it a suitable and pragmatic feature for analysis in large datasets (X. Qu et al., 2018).

2.2 Fake review detection

In the realm of online commerce, fake review detection has emerged as a pivotal component, profoundly influencing consumer behavior. These reviews, which can range from positive to negative, play a significant role in shaping perceptions of product quality and reliability, thus guiding purchasing decisions (S. Wang et al., 2023). The nature of fake reviews is such that they do not truthfully represent an authentic customer's experience, which often results in misleading potential buyers, especially given their seemingly high helpfulness ratings (Harrison-Walker & Jiang, 2023). The generation of these reviews is motivated by a variety of factors, including monetary rewards and complimentary products. In some instances, professional writers are hired by companies to craft these deceptive reviews. Additionally, there are instances of businesses engaging in reciprocal review agreements to remain undetected. The phenomenon also extends to influencers and social media personalities, who frequently participate in the orchestration of fake review campaigns. These reviews can serve different purposes: while positive fake reviews are intended to elevate a business's standing (e.g., Amazon sellers offering refunds or commissions for favorable reviews), negative ones are often directed at harming competitors' reputations (Y. Wang, Zamudio, et al., 2023). The proliferation of fake reviews, often termed review spam, poses a substantial threat to the integrity of online review systems. It is essential to secure the online review system and develop effective strategies to detect and prevent fake reviews. These fake reviews typically contain inaccurate or false information, aiming to mislead consumers into making misguided choices (Kauffmann et al., 2020). Practical implications suggest that managers should encourage honest reviews, including negative ones, to maintain credibility and

assist potential customers in their decision-making process (S. Wang et al., 2023). Fake information can affect stock price, business reputation, and also customer expectations.

2.3 Techniques for Fake Review Detection

2.3.1 NLP and ML Techniques

The identification of fake reviews falls under the purview of Natural Language Processing (NLP). NLP is capable of understanding, manipulating, and analyzing human language. NLP techniques consist of textural data preprocessing and word embeddings. By utilizing various machine learning and deep learning models, NLP methods can achieve high performance in classifying fraudulent reviews (Mridha et al., 2021).

The labelled datasets that papers in the field of fake review detection use can be obtained in three ways. In the first approach, fake reviews are manually annotated; however, human labelling is time-consuming and prone to mistakes since fake reviewers are professional users who put effort into making their reviews seem authentic (Mukherjee et al., 2012). In the second approach, fake reviews are generated and collected by crowdsourcing services and anonymous online workers. The pseudo-fake reviews are then combined with real, truthful reviews with similar lengths. Gold-standard fake reviews datasets, as known as Ott dataset, are made with this strategy in (Ott et al., n.d.) and (Ott et al., 2012) for hotels and (J. Li et al., 2014) for restaurants and doctors. However, online workers might not be experienced enough in creating convincing and professional fake reviews (Hajek & Sahut, 2022). Finally, in the third approach, fake and authentic reviews can be collected using the available filtering system embedded in platforms such as Amazon and Yelp. The filtering processes of these platforms can categorize fake and authentic reviews with high reliability. The accuracy of labels in this approach is high and large datasets can be extracted efficiently (Hajek et al., 2020).

Fake review detection is a particular case of the larger issue of deception detection. Previous research in this area has been primarily based on three approaches: review-centric, reviewer-centric, or a combination of both. For review-centric approach, linguistic features such as n-gram, word embeddings, Bag of Words (BOW), as well as language style, Part-of-Speech (POS) features, review length, and word count have been proposed. In addition to these, other proposed textual features include sentiment analysis, semantic similarity and diversity, subjectivity, a range of lexical and

syntactic features, and even more detailed features such as understandability, expressiveness, informality, complexity, level of detail, writing style, and cognition indicators. Research in this field has revealed that review texts contain valuable information, provide deep insights into customer behavior and aid consumers in making more informed decisions (Barbado et al., 2019; Goel et al., 2021; Hajek & Sahut, 2022; Shan et al., 2021). One study (L. Li et al., 2020) focused on the psychological aspects within review language by analyzing various linguistic cues such as affective, cognitive, social, and perceptual elements to distinguish between genuine and fake reviews. The study's insights highlight the nuanced ways in which language can be used to detect authenticity in online platforms. In the review-based approach, input features can be derived either from the text itself, which is referred to as textual featuring, or from the information and content of the review texts, which can be called content-based featuring (Budhi & Chiong, 2022). In the reviewer-centric approach, the focus is on the behavioral-based features rather than the content of textual reviews. Reviewer-centric features include aspects such as the user's number of written reviews, number of verified purchases, membership length, number of photos, review count, ratio of positive and negative reviews, number of votes, friends, and followers (Birim et al., 2022; Shan et al., 2021). Other features such as the time and device used to post the review have also been used to improve classification models. These features can provide additional information about the reviewer's credibility and their motivation to post a review (Abri et al., 2020; Zhang et al., 2016). Regarding reviewer's credibility, credit history has been shown to play a strong role in credibility of the written statement as well (Alghamdi et al., 2022). Choosing the right set of features is mainly based on their usefulness prediction and inference from previous similar studies. However, sensitivity analysis can also be utilized to calculate the importance of each feature (Yao et al., 2021).

In the development of fake review detection models, researchers have employed techniques such as applying sentiment analysis and leveraging unsupervised learning methods. These approaches have been proven to be effective in identifying spam comments with notable precision (Kauffmann et al., 2020). Another significant study (Rout et al., 2018) explored the use of semi-supervised learning techniques for spam detection, including co-training, Positive-Unlabeled (PU) learning, and the Expectation-Maximization (EM) algorithm. This method, focusing on review-centric textual features, achieved commendable accuracy. Concurrently, some researchers argue that non-textual features may be more effective in identifying fake reviews on online platforms (Abdulqader et al., 2022; Zhang et al., 2016). In (Barbado et al., 2019), researchers utilized a dataset from Yelp, comprising electronic

business reviews, to test both review-centric and user-centric features. The study found that while review-centric approaches, including various textual features such as bigrams and sentiment analysis, did not yield highly accurate results, the incorporation of user-centric features (which consider the personal and social information of users, along with their activity patterns) significantly enhanced the accuracy.

The debate continues in academic circles about the relative effectiveness of content-based versus behavioral-based methods for spam detection. However, several studies have demonstrated the utility of textual features in achieving satisfactory outcomes (Bathla et al., 2022; Birim et al., 2022; Budhi & Chiong, 2022; Elmurngi & Gherbi, 2017; Hajek et al., 2020; Hmoud Al-Adhaileh & Waselallah Alsaade, 2022; Liu et al., 2022). To address the limitations inherent in review-centric features, some researchers have proposed novel strategies. These include comparing writing styles for author identification and accounting for concept drift, which refers to changes in features or labels over time (Gutierrez-Espinoza et al., 2020), (Goel et al., 2021; Hajek & Sahut, 2022). In a noteworthy study (Shan et al., 2021), the researchers extracted review inconsistency features across three dimensions – rating, content, and language – and utilized them in fake review detection models. This approach aligns with the broader literature, which suggests a higher incidence of inconsistencies in fake reviews. RSI, in particular, has emerged as a significant feature in these detection models. By integrating a range of verbal and non-verbal features, researchers have been able to achieve as high as 90% accuracy in detecting fake reviews using a subset of Yelp reviews (Abdulqader et al., 2022). In a related study (Bhopale et al., 2021), researchers experimented with different combinations of n-grams to identify fake hotel reviews. They discovered that a unigram-bigram combination, implemented through a multinomial Naïve Bayes classifier, yielded the highest accuracy. Alongside n-gram analysis, this study also incorporated several other textual features, such as word count and review length, as well as basic user information, to assess the credibility of the reviewer.

Deep learning classifiers have also been extensively applied in this domain. By leveraging various textual features, including POS tags, sentiment scores, and four-grams, researchers have designed models that outperform many previous studies on fake review detection algorithms (Nagi Alsubari et al., 2022), (Hmoud Al-Adhaileh & Waselallah Alsaade, 2022), (Liu et al., 2022). These methods, notably CNN and LSTM models, have demonstrated superior efficacy compared to traditional classifiers like SVM and Naïve Bayes (Singh et al., 2023). The scarcity of labeled datasets in the realm of online fake and real reviews has prompted researchers to create their own. In a notable initiative

(Elmurngi & Gherbi, 2017), a dataset of restaurant reviews was compiled, encompassing both authentic reviews from verified users and fake reviews authored by crowd workers. This study focused on linguistic features, identifying key characteristics that distinguish fake reviews, such as their tendency to be longer and more redundant than real ones. Lexical diversity and the use of adjectives were also highlighted as significant factors. A subsequent study using the same dataset (Gutierrez-Espinoza et al., 2020) applied ensemble techniques and document embeddings to enhance the model's performance. Ensemble methods, in particular, have been shown to yield robust results. Another study (Budhi & Chiong, 2022) proposed a framework for multi-type classifier ensembles (MtCE) for detecting fake online reviews using textual-based features. These ensemble strategies have been found capable of enhancing model robustness and outperforming base classifiers when appropriately selected (Yao et al., 2021).

Emotion mining has emerged as a valuable strategy in utilizing content information for fake review detection. A study (Hajek et al., 2020) integrated emotion indicators with word embeddings and n-gram analysis into a deep learning classifier. The use of word embeddings, in conjunction with deep learning classifiers, has proved effective in addressing the complexities and abstractions arising from data sparsity and high-dimensional features. In a related study (Hajek & Sahut, 2022), researchers combined sentiment-dependent textual features with users' behavioral features to delve deeper into the true intentions of users. Despite the general assumption of consistent sentiment throughout a text, the inclusion of aspect-based polarity analysis improved the accuracy of the spam detection model. To address the limitations identified in this approach, another study (Bathla et al., 2022) extracted aspects from texts using POS tagging and then conducted sentiment analysis for these extracted aspects rather than the entire text.

The underlying topics in reviews can also serve as indicators of authenticity. In an innovative approach (Birim et al., 2022), researchers used topic distribution, applying the Latent Dirichlet Allocation (LDA) technique, as a feature in their fake review detection model. This model also incorporated other textual features, including sentiment scores, bag-of-words, and review length, as well as verified purchase data as a behavioral feature. The combination of topic features with readability features and sentiment analysis has also been demonstrated to achieve high accuracy in fake review detection (Lin et al., 2021).

Table 2 presents prior research highlighted in this section. Regarding the field of fake review detection, this table shows that previous research has been mainly centered around restaurant, hotel, and doctors (gold-standard Ott dataset) as well as Yelp datasets domains when considering online reviews.

Table 2.1 Performance of Fake Review Detection in Previous Studies

Study	Review-centric	Reviewer-centric	Classification method	Dataset	Performance
Barbado et al. (Barbado et al., 2019)	bigram, sentiments, LDA, POS, word embeddings	user personal features, #followers, #friends, #votes, #reviews, rating distribution, #photos, #tips	AdaBoost	Yelp	F-score=82
Hajek et al. (Hajek et al., 2020)	BOW n-grams, skip-gram word embeddings, lexicon-based emotion indicators	-	deep feed-forward neural network (DFFNN)	Amazon	Acc=82.8
Gutierrez-Espinoza et al. (Gutierrez-Espinoza et al., 2020)	document embedding	-	ensemble MLP	Restaurant	Acc=77.3
Abri et al. (Abri et al., 2020)	#adjectives, redundancy, pausality, lexical diversity	-	MLP	Restaurant	Acc=79.09
Goel et al. (Goel et al., 2021)	POS, unigram, bigram	rating deviation, rating, review date	NN	Yelp	Acc=79.9
Shan et al. (Shan et al., 2021)	sentiments, language style, rating-sentiment inconsistency, content inconsistency, language inconsistency	behaviorial features	Random Forest	Yelp	F-score=92.9

Birim et al. (Birim et al., 2022)	Review length, Topic distribution, Sentiment	Verified Purchase	Random Forest	Amazon	Acc=81.57
Hajek & Sahut (Hajek & Sahut, 2022)	sentiment-dependent word embedding, BOW	average rating, rating deviation, early time frame, abnormal review numbers	CNN	Yelp	F-score=83
Bathla et al. (Bathla et al., 2022)	POS, aspect-level sentiments, Doc2Vec, review length	-	CNN and LSTM	Yelp	Acc=89.5
Hmoud Al-Adhaileh & Waselallah Alsaade (Hmoud Al-Adhaileh & Waselallah Alsaade, 2022)	LIWC, Word2Vec, positive/negative words, personal pronouns, word count, rating value, authenticity, analytical thinking	-	BiLSTM	Yelp	Acc=85
Budhi & Chiong (Budhi & Chiong, 2022)	BOW, n-gram, POS	-	MtCE	Hotel	F-score=90.16
Liu et al. (Liu et al., 2022)	n-gram, word embeddings	-	CNN and BiLSTM	Hotel, Restaurant, Doctor	Acc=86.5
Yao et al. (Yao et al., 2021)	-	rating, #review, #friend, votings, active window, average rating	ensemble RF,lightgbm, catboost	Restaurant	F-score=79.65
Abdulqader et al. (Abdulqader et al., 2022)	specificity, quantity, non-immediacy, affect, uncertainty, informality, consistency	behavior deviation, source credibility	Logistic Regression	Yelp	Acc=89.73
Bhopale et al. (Bhopale et al., 2021)	n-gram, review length, words count, digits count, character count	user profile features	Multinomial NB	Hotel	Acc=88.12

Refaeli & Hajek (Refaeli & Hajek, 2021)	-	-	BERT	Hotel, Restaurant, Doctor	Acc=91
Kauffmann et al. (Kauffmann et al., 2020)	Cosine similarity measure	-	Unsupervised task (Clustering)	Amazon Product Data	Acc=85.53

2.3.2 Transformer Models

The advent of transformer-based models has opened new avenues for refining NLP tasks, including the classification of fake versus authentic reviews. Experiments using BERT and RoBERTa pretrained models, combined with GloVe embeddings, have been conducted with promising results in terms of accuracy (Refaeli & Hajek, 2021), (Mohawesh et al., 2021). This comprehensive approach highlights the multifaceted nature of fake review detection and underscores the ongoing efforts to refine and enhance methodologies in this evolving field.

2.3.2.1 BERT (Bidirectional Encoder Representations from Transformers)

BERT, introduced by Devlin et al. in 2018, is a groundbreaking model that captures bidirectional context in text. BERT's ability to understand context from both directions in a sentence made it highly effective for various NLP tasks, including question answering and text classification (Devlin et al., 2019). In the context of fake review detection, BERT has been fine-tuned on labeled datasets to identify deceptive content. A study by Refaeli and Hajek (2021) demonstrated that BERT achieved 91% accuracy on a balanced crowdsourced dataset of hotel, restaurant, and doctor reviews, and 73% accuracy on an imbalanced third-party Yelp dataset of restaurant reviews (Refaeli & Hajek, 2021).

2.3.2.2 RoBERTa (Robustly Optimized BERT Pretraining Approach)

RoBERTa, an optimized variant of BERT, was introduced by Liu et al. in 2019. The key improvements in RoBERTa include training the model longer, on more data, and with larger batch sizes. RoBERTa's enhancements make it more robust and accurate in handling various NLP tasks (Liu et al., 2019). In fake review detection, RoBERTa has been shown to perform about 7% better than state-of-the-art methods in mixed domains, achieving the highest accuracy of 91.2% on the deception dataset, indicating that RoBERTa outperforms traditional models, particularly on datasets where fake reviews are more realistic and harder to detect (Mohawesh et al., 2021). Another version of BERT is DistilBERT that aims to mitigate some of BERT's limitations, such as computational complexity. DistilBERT retains the transformer architecture but with fewer layers, making it faster while maintaining a substantial portion of BERT's accuracy. DistilBERT has also been successfully applied in the field of fake review detection, offering a balance between performance and computational efficiency (Mohawesh et al., 2021).

2.4 Feature Extraction

2.4.1 Word embeddings

Word embeddings, a sophisticated feature extraction technique, play a crucial role in the realm of NLP. These embeddings map words to a multi-dimensional vector space, wherein words with proximate vectors often share semantic and syntactic similarities. A variety of methods for computing word embeddings have emerged, among which Word2Vec, GloVe, and fastText are the most prominent. Their effectiveness is evident in a range of NLP applications, from text classification and sentiment analysis to spam detection and question-answering (Kanakaris & Karacapilidis, 2023).

The core advantage of word embeddings over traditional methods like TF-IDF and count vectorizer (CV) lies in their ability to eschew the curse of dimensionality. This is achieved by representing words as fixed-length floating-point vectors within the high-dimensional space of the vocabulary (Ren & Ji, 2017). These vectors are refined through machine learning algorithms, including neural networks and transformers (Mohawesh et al., 2021) which learn from the context provided by surrounding words.

As a result, the dense vectors that emerge enable words with analogous meanings to cluster together spatially.

The creation of word embeddings can occur concurrently with the main task using an embedding layer or through pretraining on a substantial standard corpus, followed by application to downstream tasks. Renowned pre-trained word embeddings, such as BERT, fastText, GloVe, and Word2Vec, encapsulate vast amounts of domain or language-specific knowledge acquired during training, thereby reflecting statistical word correlations within those specific contexts (Chollet, 2018).

Past studies show that hybrid deep learning approaches such as BiGRU can achieve good results when using pretrained word embedding (Rao et al., 2023). GloVe is a global log-bilinear regression model for the learning of word embeddings that has superior performance in terms of word similarity and word analogy. GloVe is based on global word co-occurrence statistics and can also capture meaningful local information, akin to prediction-based methods like Word2Vec (Pennington et al., 2014). Since GloVe is a pretrained embedding, it can improve both efficiency and performance and capture rich semantic and syntactic information due to its training on large corpora. On the other hand, Word2Vec learns patterns by focusing on local context and the window of words around a target word. Word2Vec either uses Continuous Bag of Word (CBOW) or Skip-Gram approaches. In CBOW, the model predicts a word based on its context, whereas, in Skip-Gram, it predicts the context given a word (Goldberg & Levy, 2014).

2.4.2 Document embedding

Document embedding, as described in (Y. Qu et al., 2022) involves the concatenation of sentence embeddings or is derived from sentence embeddings through a neural model. Document-level representation is another method of feature representation that rather than focusing only on discrete features considers document-level semantics (Bathla et al., 2022). Doc2vec is an extension of Word2vec in which embedded vectors are documents. Document representation has shown to be effective in detecting deceptive opinions (Bathla et al., 2022; Gutierrez-Espinoza et al., 2020; Ren & Ji, 2017). (Ren & Ji, 2017) represents text using vector representations for sentences and documents. The generated document composition is then used to capture semantic relations between sentences and detect spam reviews.

2.5 Limitations of Previous Works

While significant advances have been made in the field of fake review detection, several limitations exist in previous research. Many studies have focused either on review-centric or reviewer-centric features but have often neglected the integration of the two approaches. This creates gaps in detection effectiveness, particularly when distinguishing subtle differences in textual content or reviewer behavior. Additionally, most prior models rely heavily on features like review length, sentiment, and simple word frequency counts, which may overlook more nuanced indicators of deception such as RSI.

Furthermore, many existing works have primarily employed traditional machine learning techniques such as Naive Bayes, Support Vector Machines (SVMs), and Decision Trees, which, although useful, may not capture complex relationships in the data. These models lack the capacity to fully exploit deep learning's ability to learn hierarchical representations of text, which is crucial for improving the accuracy of fake review detection.

Previous studies have also focused on limited datasets such as Yelp and Amazon, often generating datasets through crowdsourcing or relying on filtering systems that may introduce bias. This restricts the generalizability of the models across different platforms and content types. Additionally, few studies have explored the role of pre-trained embeddings in conjunction with textual inconsistency metrics.

This study seeks to address some of the gaps identified in previous works by introducing novel approaches to fake review detection. Key contributions include:

- Integration of Review-Centric and Reviewer-Centric Features: By leveraging Rating-Sentiment Inconsistency (RSI) alongside pre-trained word embeddings and document embeddings, this study bridges the gap between review-centric and reviewer-centric features, addressing the need for more holistic detection models.
- Novel Combination of RSI with Pre-trained Embeddings: Few studies have explored the use of inconsistency metrics like RSI combined with powerful pre-trained word and document embeddings. This study introduces two novel combination methods that improve model accuracy, addressing the gap in feature engineering for fake review detection.
- Improved Generalizability: Unlike previous works that rely on limited datasets or crowdsourced data, this study proposes methods that can generalize across different platforms

by utilizing widely accessible features (review text and ratings) and combining them with deep learning techniques. This approach reduces reliance on manually labeled or filtered datasets.

- Higher Detection Accuracy: The integration of RSI with advanced embedding techniques (e.g., GloVe, Word2Vec, Doc2Vec) has demonstrated a substantial improvement in classification accuracy, outperforming traditional models by up to 3.07%, thus highlighting the practical implications of this approach in real-world applications.

Chapter 3 Methodology

We propose two frameworks for combining RSI with review representation as shown in Fig. 3 and Fig. 4, exhibited in 3.2 and 3.3. In the first method, we concatenate the normalized values of RSI with the review representations learned from pretrained word embeddings as separate inputs. This approach assists in considering the global influence of inconsistency in the model input and is identified as C1. Drawing inspiration from (Du et al., 2019; X. Qu et al., 2018), the second method treats normalized RSI values as new vocabulary items. These values are then trained as individual words for each review, alongside document embeddings, to learn the content representations. It is essential to convert the inconsistency information into a vector matching the review's dimensionality in the document embedding layer. The rationale for choosing document embeddings (doc2vec) over word embeddings in this method stems from the fact that RSI values are review-specific and can be integrated within document embeddings, which also pertain to individual reviews. In contrast, combining RSI values with word-specific word embeddings in a single matrix representation is not feasible. We refer to this approach as C2. The RSI and proposed methodologies C1 and C2 will be detailed in 3.1, 3.2, and 3.3.

To test the performance of our proposed methodology and framework, we are also testing baseline methods, including BiGRU, CNN, RNN, LSTM, and BiLSTM. These baseline models will help benchmark the effectiveness of our approach against existing neural network architectures widely used in fake review detection tasks. The generic framework of this study is depicted in Figure 1.

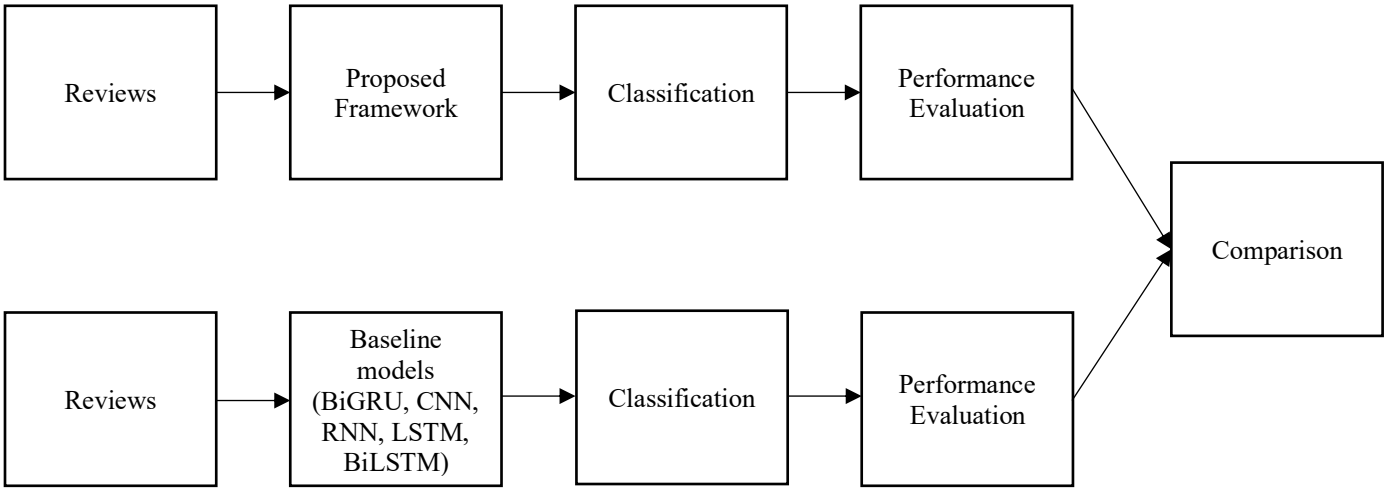

Figure 3.1 Research Framework

3.1 Rating-sentiment Inconsistency (RSI)

Prior to calculating the RSI, we subjected the sentiment scores to min-max normalization to linearly adjust the values to a uniform scale ranging from 1 to 5. This normalization ensures that sentiment scores are directly comparable to star ratings, thereby mitigating the influence of outliers on the analysis. The normalization is implemented in equation (1) as follow:

$$NormalizedSentiment_i = \frac{sentiment_i - \min(Sentiments)}{\max(Sentiments) - \min(Sentiments)} \quad (1)$$

Then for each review i in our analysis, we quantified the RSI by determining the absolute deviation between the normalized sentiment of the review and its star rating, as delineated in equation (2).

$$RSI_i = |NormalizedSentiment_i - Rating_i| \quad (2)$$

In a similar study, the standard deviation and mean of sentiments and ratings for identical products were utilized to obtain normalized values. However, due to the unique product review distribution within our Amazon dataset, we instead used the aggregate minimum and maximum sentiment scores

to normalize individual review sentiment scores. Consequently, the resultant RSI scores span from 0 to 3.6, where a lower score signifies minimal inconsistency, and a higher score denotes greater inconsistency.

For the extraction of sentiment scores from reviews, we employed SentiWordNet, a rule-based lexical resource renowned for its accuracy in opinion mining. We calculated the overall sentiment score for each review by summing the sentiment values and normalizing by the length of the review. The preprocessing of the review text was meticulously conducted through several steps: elimination of special characters, tokenization, conversion to lowercase, removal of stop words, and application of stemming and lemmatization techniques. Such preprocessing is imperative to exclude extraneous data and minimize the impact of noise during the training phase (Mridha et al., 2021).

Figure 2 presents a boxplot visualization of the RSI scores segregated by various rating categories. The inconsistency manifests variably across different rating levels, with outliers accentuating the potential for discrepancy within each category. Notably, reviews with extreme ratings (1 and 5) tend to show higher levels of inconsistency, characterized by elevated median values and broader interquartile ranges, while reviews with moderate ratings exhibit a reduction in inconsistency.

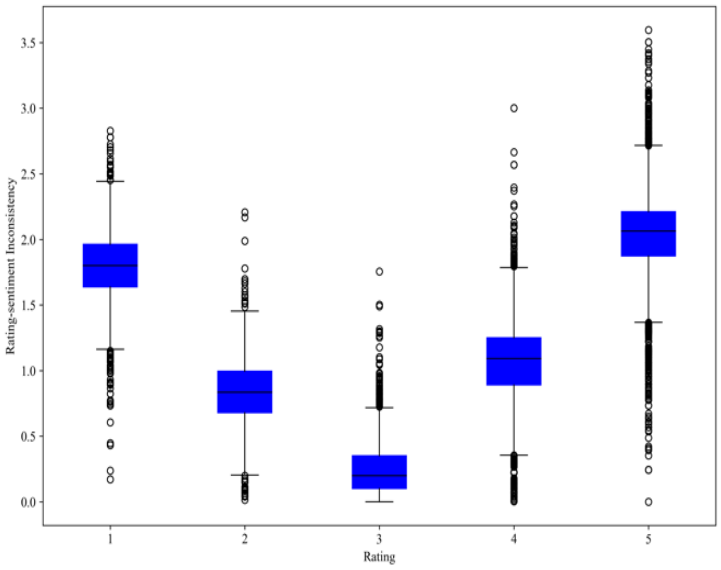

Figure 3.2 Boxplots of rating-sentiment inconsistency distribution based on rating

3.2 C1 Architecture

In the C1 approach, review texts undergo a tokenization process and are subsequently padded to a fixed length of 100 tokens. The average length of the reviews in this dataset is 32 tokens, with the longest review extending to 1024 tokens. While we experimented with longer padding lengths, we observed limited efficacy, and thus concluded that a length of 100 tokens is a suitable choice. This length also allows the model to process reviews efficiently, balancing computational complexity with performance. Additionally, 100 tokens aligns with standard practices in text classification tasks, making it a practical choice for model efficiency and generalization. For our embedding dimension in both GloVe and Word2Vec experiments, we selected 300 as the optimal dimension. The architectural framework of our model is defined using the Keras framework. It is structured around two primary input layers: one for the tokenized text data and the other for the RSI values. The tokenized and padded text data is subsequently passed through an embedding layer, which leverages pre-trained word embeddings from an embeddings matrix to map individual words to their corresponding vector representations.Following the embedding layer, the model incorporates a BiGRU (Bidirectional Gated Recurrent Unit, see sub-section 3.5) layer to process the embedded text data and capture contextual information effectively. The output of the BiGRU layer is concatenated with the input representing RSIs. This concatenation operation is then passed through a dense layer employing a Sigmoid activation function to generate a probability value within the range of 0 to 1 for each output node.

The concatenation layer operates on a list of tensors as input, merging them along a specified axis to create a unified output tensor. In this context, each row within the resulting tensor corresponds to an individual sample within the batch, facilitating the fusion of information from both the BiGRU layer and the inconsistency data sources.

The GloVe embeddings used were trained on Wikipedia and Gigaword (6B tokens, 400k vocab, uncased, 300d vectors) (*GloVe Embeddings*, 2015). To load the pre-trained word embeddings from the GloVe file, the code initializes an embedding matrix with zeros, and then iterates through the GloVe embeddings. This process assigns a vector representation to its corresponding token, within the specified maximum number of features, in the embedding matrix. The resulting embedding matrix is fed later into the embedding layer of the deep learning classifier.

The Google News Word2Vec embeddings are loaded using *KeyedVectors.load_word2vec_format.* These embeddings are trained using the Skip-Gram model on a large dataset from Google News, consisting of about 100 billion words, resulting in word vectors for about 3 million words and phrases.

Our approach to parameter tuning involves the application of a grid search (Liashchynskyi & Liashchynskyi, 2019) optimization method to identify the most optimal parameter configurations. Grid search conducts a complete exploration across a specified subset of the training algorithm's hyperparameter space. Given the manageable size of our hyperparameter search space and the critical need for precision in the detection of fake reviews, grid search optimization approach have been employed in this study (Liashchynskyi & Liashchynskyi, 2019).

For the experiment with GloVe embeddings, the model employs a series of Bidirectional GRU layers, starting with 128 units and progressively halving the number of units in subsequent layers down to 64, 32, and finally 16 units, with the aim to capture complex dependencies in the data from both directions. After each GRU layer, a dropout rate of 0.3 is introduced to mitigate overfitting by randomly omitting a fraction of the neurons during training. The network then transitions to a dense layer consisting of 64 units with a ReLU activation function to introduce non-linearity, followed by an output layer with a single unit employing a sigmoid activation function for binary classification. Adam optimizer is utilized with a learning rate set at 0.001 to optimize the binary cross-entropy loss function.

For the experiment with Word2Vec, the text is processed through a series of bidirectional GRU layers, each followed by a dropout layer with a rate of 0.2 to prevent overfitting. The model culminates in a dense layer with a sigmoid activation for binary classification. It is optimized with the Adam algorithm over 10 epochs and in batches of 64. This configuration, including fixed embedding dimensions, GRU units (64, reducing by half in subsequent layers until 16), and the dropout rate, forms the backbone of the model's hyperparameters setup. To help explain how the data flows through the architecture and how it is processed at each step, corresponding equations and algorithm are outlined below:

Steps	**Operation/Equation**
Step 1: Tokenize and Pad Text	Input review texts are tokenized and padded to 100 tokens.
Step 2: Word Embedding	$X_{embed} = W_{embed} \cdot X_{tokens}$ (Word2Vec or GloVe embeddings with dimension 300)

Step 3: BiGRU Layer	Forward GRU: $\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}})$
	Backward GRU: $\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t+1}})$
	Combined output: $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$
Step 4: Concatenation	Concatenate h_t (BiGRU output) with RSI values: $Z = [h_t, RSI]$
Step 5: Dense Layer (ReLU Activation)	$A = ReLU(W_d.Z + b_d)$
Step 6: Output Layer (Sigmoid Activation)	$\hat{y} = \sigma(W_o.A + b_o)$
	(Binary classification output)
Step 7: Loss Function	Binary Cross-Entropy Loss:
	$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i}))$

Where:

- X_{embed} is the embedded representation of the tokenized text.
- W_{embed} is the pretrained word embeddings matrix (GloVe/Word2Vec).
- X_{tokens} is the matrix of tokenized input text.
- h_t is the concatenated hidden state from forward and backward GRU cells.
- x_t is the tokenized input at a time step t.
- W_d is the weight matrix of the dense layer.
- b_d is the bias term of the dense layer.
- $\hat{y}$ is the predicted output (probability of the review being fake).
- W_o is the weight matrix of the output layer.
- y_i is the true label for review i.
- $\hat{y_i}$ is the predicted probability for review i.

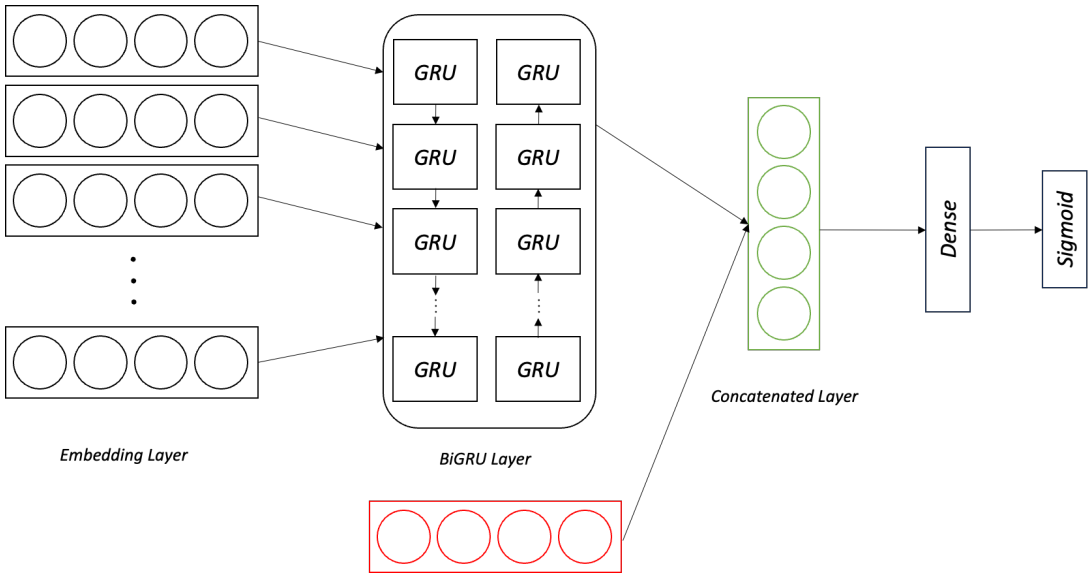

Figure 3.3 C1 Method Architecture (Red box indicates RSI vector and green box denotes the concatenated layer)

3.3 C2 Architecture

In the C2 approach, we employ a Doc2Vec model to process tokenized reviews. Each review is tagged with a unique identifier, and the model is trained with specific configuration settings, including a vector size of 300, a window size of 5, a minimum count of 1, and a total of 10 training epochs. Our code iteratively processes these tagged reviews, generating sentence-level embeddings for each review utilizing the trained Doc2Vec model. These resulting embeddings are then compiled into a list. This list of sentence embeddings is subsequently combined with the RSI values, effectively creating a unified feature matrix with a resulting shape of (21,000, 301). In this configuration, the number 21,000 represents the total count of reviews, while 301 corresponds to the number of features contained within the matrix. These features encompass 300 dimensions from the Doc2Vec embeddings, and an additional 1 feature derived from the inconsistency values.

This resulting feature matrix, which encompasses both the embeddings and inconsistency values, undergoes a reshaping process to form a 3D input structure. This transformation is necessary to ensure compatibility with the Keras model equipped with a BiGRU classifier. More specifically, our classifier architecture consists of two BiGRU layers, a dropout layer, and a dense layer tailored for binary classification. The first BiGRU layer consists of 128 units, followed by the second BiGRU

layer with 64 units. We introduce a dropout layer with a rate of 0.2 to mitigate overfitting. The activation function applied in the dense layer is Sigmoid, facilitating binary classification. We used grid search method as the selection method for these hyperparameters. Corresponding equations and algorithm are outlined below:

Steps	**Operation/Equation**
Step 1: Tokenize and Tag Reviews	Each review is tokenized and tagged with a unique identifier.
Step 2: Doc2Vec Embedding	Sentence-level embedding from Doc2Vec: $v_i = Doc2Vec(r_i)$ (300-dimension vector for each review)
Step 3: Concatenate Doc2Vec with RSI	Combine Doc2Vec embeddings with RSI values: $Z_i = [v_i, RSI_i]$
Step 4: Reshape Input for BiGRU	Reshape the input to 3D format: $X = reshape(Z, (n_{samples}, t_{timesteps}, f_{features}))$
Step 5: BiGRU Layer	Process input through BiGRU: Forward GRU: $\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}})$ Backward GRU: $\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t+1}})$ Combined output: $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$
Step 6: Dense Layer (ReLU Activation)	$A = ReLU(W_d . h_t + b_d)$
Step 7: Output Layer (Sigmoid Activation)	$\hat{y} = \sigma(W_o . A + b_o)$ (Binary classification output)
Step 8: Loss Function	Binary Cross-Entropy Loss: $L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i}))$

Where:

- v_i is the vector representation of review i generated by Doc2Vec.
- r_i is the tokenized review i.
- X is the reshaped 3D matrix.
- n_{sample} is the number of reviews.
- $t_{timesteps}$ is the steps (1 in this case since it's a sentence-level embedding).
- $f_{features}$ is the number of features (301:300 from Doc2Vec and 1 from RSI).

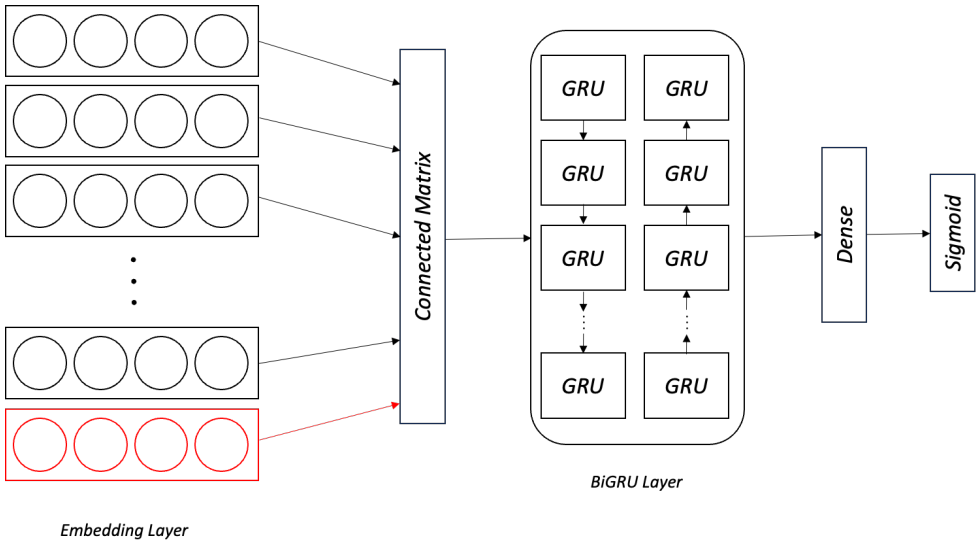

Figure 3.4 C2 Method Architecture (Red box indicates RSI vector)

3.4 Motivation of the Proposed Framework

The motivation for developing a combined model integrating review inconsistency features with deep learning techniques for fake review detection stems from several key observations and objectives.

- Traditional detection methods, which often rely on either content-based features or reviewer behavioral patterns, have shown limitations in achieving high accuracy and robustness. By leveraging the complementary strengths of various features, a combined approach promises to enhance detection capabilities.
- The integration of Rating-Sentiment Inconsistency (RSI) as a feature is motivated by its potential to capture subtle cues of deception that are often overlooked by other methods. RSI highlights discrepancies between the sentiment expressed in the review text and the associated numerical rating, providing a nuanced indicator of possible fraudulent intent. Prior studies have demonstrated that such inconsistencies are prevalent in fake reviews, making RSI a valuable addition to the detection framework.
- The application of pre-trained word embeddings (such as GloVe and Word2Vec) and document embeddings (such as Doc2Vec) provides a robust representation of review texts. These embeddings encapsulate semantic and syntactic nuances, allowing the model to better

understand and classify the content. Combining these embeddings with RSI ensures that the model benefits from both advanced text representation techniques and specialized inconsistency metrics.

- The choice of deep learning architectures, particularly Bidirectional Gated Recurrent Units (BiGRUs), is motivated by their ability to capture contextual dependencies in sequential data. BiGRUs process text in both forward and backward directions, thereby enhancing the model's ability to detect contextually relevant patterns indicative of fake reviews.

3.5 Bidirectional Gated Recurrent Unit (BiGRU) Classifier

GRUs (Gated Recurrent Unit) are a more streamlined variant of LSTMs, optimized for handling temporal data without the need for separate memory cells. They utilize two gates—update and reset—to modulate the flow of information within the network. GRUs excel in managing long-distance dependencies in sequence data by adjusting the influence of word length, which aids in preventing overfitting on small sample sizes and contributes to greater model stability (Luvembe et al., 2023). The Bidirectional Gated Recurrent Unit (BiGRU) is an advanced deep learning model for processing sequences, designed to capture information from both past and future states to enhance context understanding. This bidirectional approach, which employs a pair of GRU networks operating in opposite directions, offers a more nuanced comprehension of text sequences compared to unidirectional GRUs and RNNs, making it particularly adept for detecting fake reviews (Badri et al., 2022). The BiGRU's strength lies in its ability to concurrently process features from both directions of the data sequence, thus efficiently gathering contextual details that are essential for robust predictive analytics. BiGRU is more effective than GRU since it combines two GRUs in two directions (Luvembe et al., 2023).

In this study, for the regular BiGRU experiment serving as the baseline, the BiGRU layer comprises 128 units and includes a kernel regularizer. A dropout rate of 0.2 is applied, the first dense layer consists of 64 units, and the second layer consists of 1 unit. The LeakyReLU activation function for the first dense layer and Sigmoid for the second layer are utilized, with batch normalization applied after dropout and before the final dense layer. L2 regularization is employed on both the GRU and the first dense layer. The model uses the Adam optimizer with a learning rate of 0.001. A learning rate

scheduler, specifically ReduceLROnPlateau, is utilized for the BiGRU model. These hyperparameters are selected based on grid search method.

Input Layer: The input is tokenized and padded to 100 tokens.

$$X_{tokens} = tokenize(r_i) \quad (3)$$

Where r_i is the input review.

Embedding Layer: The tokenized input is passed through an embedding layer that maps the tokens to 300-dimensional vectors using pretrained word embeddings:

$$X_{embed} = W_{embed} . X_{tokens} \quad (4)$$

Where X_{embed} is the embedded matrix and W_{embed} is the pretrained embeddings matrix.

BiGRU Layer: The embedded text is processed through a bidirectional GRUwith 128 units (using update and reset gates):

Forward GRU:

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \quad (5)$$

Backward GRU:

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t+1}}) \quad (6)$$

Concatenated Output:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (7)$$

Dense Layer: The BiGRU output is passed through a fully connected dense layer with 64 units and a LeakyReLU activation function:

$$A = LeakyReLU(W_d . h_t + b_d) \quad (8)$$

Where:

- W_d is the weight matrix of the dense layer.
- b_d is the bias term.
- LeakyReLU introduces a small slope for negative inputs.

Batch Normalization Layer: Batch normalization is applied after the dropout and before the final dense layer to stabilize learning.

Output Layer: The output layer is a single unit with a Sigmoid activation to predict the probability of the review being fake:

$$\hat{y} = \sigma(W_o . A + b_o) \quad (9)$$

Optimizer and Learning Rate:

The Adam optimizer is applied with a learning rate of 0.001, and the learning rate scheduler ReduceLROnPlateau adjusts the learning rate during training based on validation performance.

Loss Function (Binary Cross-Entropy):

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} (y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \quad (10)$$

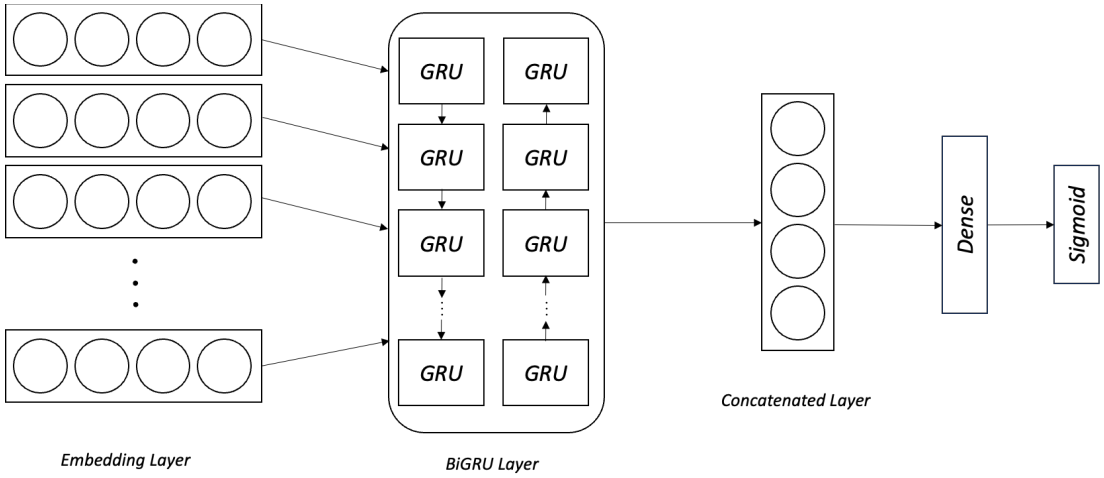

Figure 3.5 BiGRU Architecture

3.6 Baseline experiments

This paper proposes a model for detecting fake reviews by combining RSI with embedding, so as to further illustrate the performance of the model compared with other models. In baseline experiments, the only input to the model is padded sequences as opposed to our proposed model in which we added a second input to include RSI feature.

(1) Convolutional Neural Network (CNN): CNN (LeCun, 1989) is the most popular and utilized type of deep learning network. A CNN architecture typically consists of multiple layers, including convolutional layers, pooling layers, activation functions, fully connected layers, and loss functions. Each layer plays a specific role, with convolutional layers responsible for feature extraction through filters, pooling layers reducing the dimensionality of the data, activation functions introducing non-linearities into the network, fully connected layers making decisions based on the extracted features, and loss functions measuring the performance of the network

(Alzubaidi et al., 2021). CNN is also the most widely used deep learning classification method in fake review detection models (Thompson et al., 2022). Local information and semantic relationship within review texts can be captured well in the convolution operation (Liu et al., 2022). To construct these CNN in this study, Conv1D layers are configured with 128 filters and a kernel size of 5. MaxPooling1D layers have a pool size of 2. The first dense layer consists of 64 units with a ReLU activation function. The output dense layer has a single unit with a sigmoid activation function. The Adam optimizer, set with a learning rate of 0.001, is employed across the models.

Input Layer: Tokenized and padded text sequences.

Embedding Layer: Pre-trained word embeddings.

Conv1D Layer: 128 filters with a kernel size of 5. The equation for convolution operation is as below:

$$Z = (X * W) + b \tag{11}$$

Where Z is the output of the convolutional layer, X is the input matrix, and W is the filter matrix.

MaxPooling1D Layer: Pool size of 2.

$$P = \max(Z_{window}) \tag{12}$$

Dense Layer: Fully connected layer with 64 units and a ReLU activation function.

$$A = ReLU(W_d.P + b_d) \tag{13}$$

Output Layer: A single unit with Sigmoid activation.

$$\hat{y} = \sigma(W_o . A + b_o) \quad (14)$$

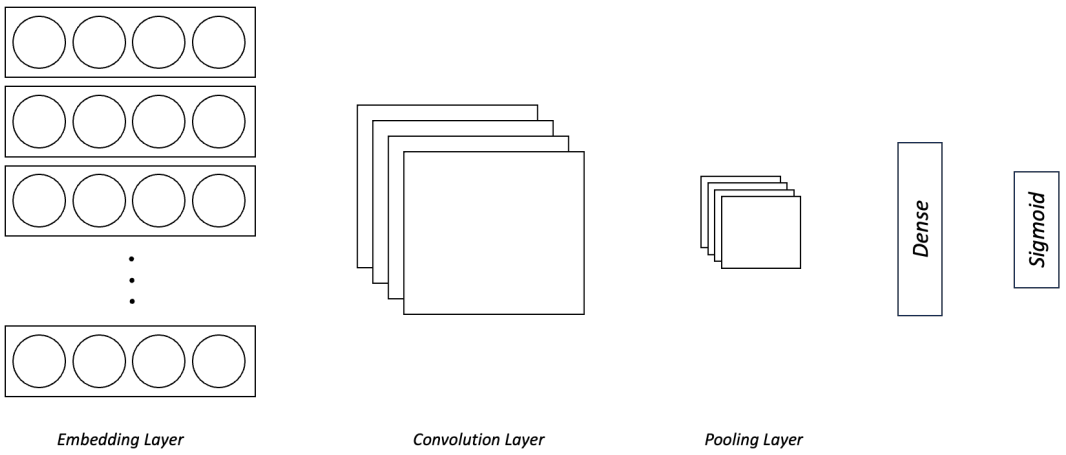

Figure 3.6 CNN Architecture

(2) Recurrent Neural Network (RNN): RNNs (Rumelhart et al., 1986) are a type of deep neural network architecture designed primarily for handling sequential data. Unlike traditional neural networks, which assume independence between inputs, RNNs can maintain a form of internal state or memory that allows them to incorporate information from previous elements in a sequence into the current context. This makes RNNs particularly well-suited for tasks in natural language processing (NLP), where understanding the sequence and context of words is crucial (Yin et al., 2017). RNNs are widely used in time-series based data and various NLP tasks such as phishing and spam email detection. By using memory feature, all computational data are stored and the output in the sequence is dependent on the preceding parts (Doshi et al., 2023). In this study, for the RNN construction, a dropout rate of 0.2 is applied, and GRU layers are set up with 128 units. The initial dense layer is equipped with 64 units and utilizes a ReLU activation function. The final dense layer features a single unit with a sigmoid activation function. Throughout the models, the Adam optimizer is used, configured with a learning rate of 0.001.

Input Layer: Tokenized and padded text sequences.

GRU Layer: 128 units.

$$h_t = GRU(x_t, h_{t-1}) \quad (15)$$

Dropout Layer: Dropout rate of 0.2.

Dense Layer: Fully connected layer with 64 units and a ReLU activation function.

$$A = ReLU(W_d . h_t + b_d) \quad (16)$$

Output Layer: A single unit with Sigmoid activation.

$$\hat{y} = \sigma(W_o . A + b_o) \quad (17)$$

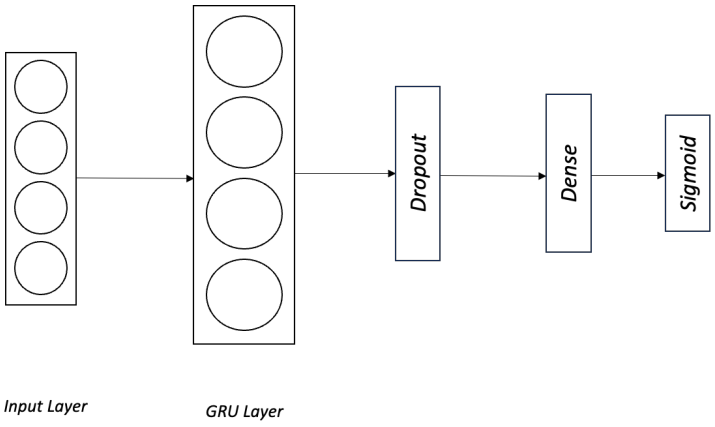

Figure 3.7 RNN Architecture

(3) Long Short-Term Memory (LSTM): LSTM (Hochreiter & Schmidhuber, 1997) networks are a type of Recurrent Neural Network (RNN) specifically designed to avoid the long-term dependency problem inherent to traditional RNNs. LSTM is appropriate for handling sequential data with long-term dependencies including textual data and text classification tasks. They achieve this through a complex architecture that includes memory cells and gates (Sherstinsky, 2020). This architecture is designed to remember information for long periods, effectively overcoming the vanishing gradient problem, a limitation in RNNs (Bathla et al., 2022). In the LSTM construction for this study, a dropout rate of 0.2 is implemented, and LSTM layers are configured with 64 units. The first dense layer comes with 32 units and employs a ReLU activation function. The concluding dense layer contains a single unit with a sigmoid activation function. The Adam optimizer, with a learning rate of 0.001, is utilized across the models.

Input Layer: Tokenized and padded text sequences.

LSTM Layer: 64 units.

$$C_t = f_t . C_{t-1} + i_t . \tilde{C}_t \qquad (18)$$

Dense Layer: 32 units with ReLU activation.

$$A = ReLU(W_d . C_t + b_d) \quad (19)$$

Output Layer: A single unit with Sigmoid activation.

$$\hat{y} = \sigma(W_o . A + b_o) \quad (20)$$

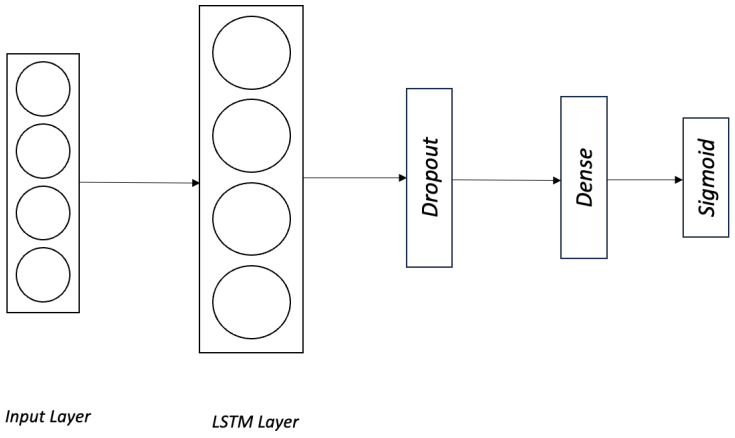

Figure 3.8 LSTM Architecture

(4) Bidirectional Long Short-Term Memory (BiLSTM): BiLSTMs (Schuster & Paliwal, 1997) extend the LSTM framework by processing data in both forward and backward directions, effectively doubling the architecture. A BiLSTM consists of two LSTM layers that run parallel to each other. For instance, a given layer processes the input sequence as it is, from the beginning to the end (forward pass), while the other processes a reversed copy of the input sequence, from the end to the beginning (backward pass). The outputs from both layers are combined at each time step, to produce a single output that reflects the context from both directions. Their ability to incorporate both past and future context into the decision-making process at any point in the sequence makes them beneficial for tasks where understanding the entire sequence is crucial for accurate predictions or classifications, such as in NLP tasks (Hmoud Al-Adhaileh & Waselallah Alsaade, 2022). In this study, for the BiLSTM setup, a dropout rate of 0.2 is applied, with LSTM

layers set to 128 units. The initial dense layer is designed with 64 units, using a ReLU activation function. The final dense layer is comprised of a single unit, equipped with a sigmoid activation function. Throughout these configurations, the Adam optimizer is employed, set with a learning rate of 0.001.

Input Layer: Tokenized and padded text sequences.

BiLSTM Layer: 128 units.

Forward LSTM:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}) \quad (21)$$

Backward LSTM:

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}}) \quad (22)$$

Concatenated Output:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (23)$$

Dense Layer: 64 units with ReLU activation.

$$A = ReLU(W_d . h_t + b_d) \quad (24)$$

Output Layer: A single unit with Sigmoid activation.

$$\hat{y} = \sigma(W_o . A + b_o) \quad (25)$$

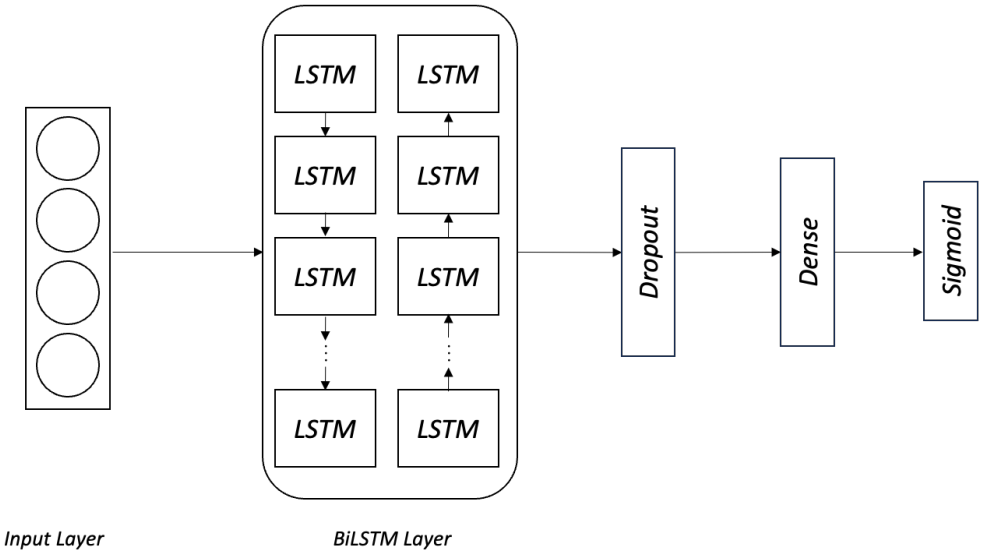

Figure 3.9 BiLSTM Architecture

Chapter 4 Experimental Setup

4.1 Dataset: Amazon dataset

Constructing models for detecting fake reviews poses a significant challenge due to the difficulty in obtaining accurately labeled data. Previous research often relied on pseudo-fake reviews, which were typically generated and labeled manually by individuals recruited for the task. However, the reliability of human-labeled data is questionable, as distinguishing fake reviews from genuine ones is challenging; fake reviews are crafted to emulate the tone and style of legitimate reviews closely. Moreover, crowd workers do not share the same motivations or psychological profiles as the original authors of fake reviews, which can lead to discrepancies in the data. Consequently, models built on such pseudo-fake reviews might not perform effectively when applied to real-world scenarios (Shan et al., 2021).

To mitigate these issues, our study leveraged authentic fake and real reviews directly sourced from Amazon (Amazon's standard fake product reviews dataset). This approach ensures a more accurate representation of the types of reviews that the model will encounter in practical applications. The selection of an ample and representative dataset is also pivotal for attaining high levels of accuracy in classification. Our research utilized a well-regarded dataset from Amazon, which comprises 21,000 reviews (*Amazon Reviews Dataset*, 2018). Amazon's own filtering system labels half of these reviews

as fake, ensuring a balanced composition. Table 3 presents five randomly selected examples of fake and real reviews. The dataset encompasses various details of each review, such as the label, rating, review and product titles, product ID, category, and verification status of the purchase. It does not, however, include data on the users' behavior or profiles. With an average review rating of 4.12 stars and representation across 30 diverse product categories, the dataset's focus on content-based features allows us to concentrate on analyzing content and linguistic attributes. Our goal is to achieve credible classification performance primarily through these features.

Table 4.1 Examples of Fake and Real Reviews

Fake	Rating	Real	Rating
When least you think so, this product will save the day. Just keep it around just in case you need it for something.	4	Great little item for the price. Good for occasional use when you are in the boonies and packing light.	5
This kettle is a gift for my daughter. She said that it is easy to use. She likes the design of it, so nice and neat.	4	Very nice! Smooth, clean sound. Good deal. Thanks. These headphones are a pleasant addition to my stereo system. I will be back!!!	4
I have had my camera for about 2 weeks now. The pictures are great, it is so easy to use. I love it!!	3	I was looking for a narrow piece with storage in a small bathroom. Very nice looking and fit well where I needed it	5
The sides of the case didn't match my phone at all but the real issue is when the tan color became blackish in some parts. Looks unpleasant and can't use like that.	1	I was looking for a winter hat with a large brim and would have liked the one pictured. The one that i received was absolutely NOTHING like the picture. The one I received had no brim at all and had a leather strip across the front in brown. The hat was black. That was the only thing that was accurate. I wouldn't pay the shipping for the return to China so I just gave it away.	1
After using lipogaine minox for a few months decided to give the shampoo a shot. Absolutely hate IT! MADE MY HAIR shed tremendously, and it smells awful. I would advise to stay away!!! ALL REVIEWS on this shampoo seem FAKE.	1	Bought 3. Bought in a bass on second cast, nice action. The second pole broke trying to bring in a medium bluegill. Have yet to use third pole. Reels are nice too.	3

4.2 Word Cloud Visualization

To visually represent the linguistic characteristics of fake and real reviews, we generated word clouds for each category. Word clouds highlight the most frequently occurring words, with word size indicating frequency. In the word cloud for fake reviews, we see a prevalence of positive and persuasive language, with common words like 'love,' 'great,' 'use,' and 'good.' This language often conveys strong positive sentiment, which aligns with the purpose of many fake reviews: to manipulate consumer perception by exaggerating product qualities. In contrast, the word cloud for real reviews reveals a more balanced and varied use of language. Although words like 'love' and 'great' are still prominent, there's a notable presence of more neutral and descriptive terms like 'work,' 'time,' 'make,' and 'need.' This variety suggests a more authentic and detailed discussion of product features and performance.

These word clouds offer a visual starting point for understanding language patterns associated with authenticity. They highlight potential features for further analysis, such as the use of exaggerated positive language in fake reviews, which our model can use to improve detection accuracy. By examining these patterns, we gain insights into how linguistic tendencies differ in real and fake reviews, which is useful for feature engineering and model training.

Figure 4.1 Top Words for Fake Reviews

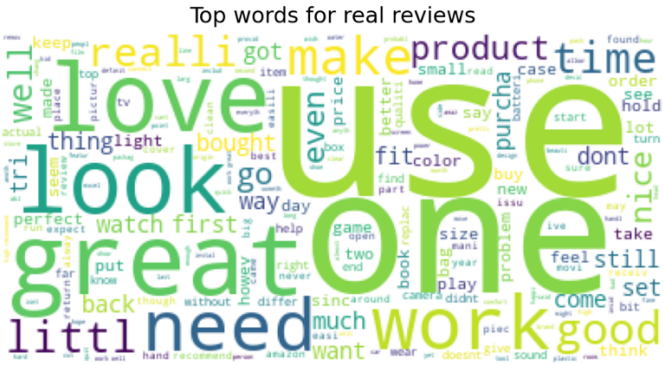

Figure 4.2 Top Words for Real Reviews

4.3 Evaluation

We performed 5-fold cross-validation, using the *StratifiedKFold Python library*, in all experiments and reported the average performance of the 5 iterations. In the first set of experiments, we investigated the effect of the RSI feature on the performance of classification models. Table 3 contains the performance results for each experiment based on accuracy, F-score, and Area Under the ROC (Receiver Operating Characteristics) Curve (AUC). To find the best set of hyperparameters for each experiment, we performed grid searches. We tested all experiments with 10 epochs, 300 embedding dimensions, and an adaptive scheduler with Keras' *ReduceLROnPlateau* callback to adjust the learning rate during training based on the validation loss.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{26}$$

$$Precision = \frac{TP}{TP + FP} \tag{27}$$

$$Recall = \frac{TP}{TP + FN} \tag{28}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (29)$$

Accuracy is the percentage of true positives and negatives divided by all predictions, precision is the number of actual positive predictions divided by the total number of positive results, F-score is the combination of precision and recall, and AUC is the probability that a classifier ranks a random real review higher than a random fake review (Hajek et al., 2020). F-score is usually used for evaluation when the dataset is not balanced (Mridha et al., 2021).

Chapter 5 Analyses and Results

5.1 Presence of rating-sentiment inconsistency

To verify the presence of RSI and in the pursuit of robust fake review detection model, we are justifying the incorporation of RSI feature by a comprehensive examination of various statistical analyses. The convergence of findings from hypotheses testing, effect size measurement, non-parametric testing, and ANCOVA collectively strengthens the rationale for the inclusion of RSI in the model.

The Mann-Whitney U test, a non-parametric test comparing the distribution of RSI between fake and real reviews, yielded a statistically significant difference (p-value < 0.001). We also performed ANCOVA, treating RSI as the independent feature and the label as a covariate. The results showed significant difference between the effect of reviews labels (fake vs. real) on RSI, based on the F-statistics of 8.684 and the associated p-value of 0.0032. The consistency in the direction of significance across multiple statistical approaches further supports the importance of RSI and its discriminative power in identifying fake reviews.

5.2 Assessing Impact of Usefulness of the Inconsistency Feature

Table 4 shows the performance results for each experiment. The empirical results from our experiments elucidate the significant contribution of RSI to the accuracy of fake review detection. Notably, the incorporation of this feature with GloVe embeddings resulted in a remarkable accuracy

improvement of 3.07%, achieving a final accuracy of 90.21%, an AUC of 92.58%, and an F-score of 90.07%. This enhancement was similarly observed with Word2Vec embeddings, albeit to a lesser extent, with an accuracy improvement of 0.67%. These findings are pivotal, evidencing the robustness of our proposed frameworks, particularly when combined with GloVe embeddings.

The Doc2Vec embedding also demonstrated an appreciable performance enhancement. When integrated with RSI, we witnessed a 1.55% increase in accuracy, substantiating the efficacy of inconsistency metrics in this domain. The comparative analysis with baseline classifiers as shown in table 5 —including CNN, RNN, LSTM, and BiLSTM—further accentuated the superior performance of our BiGRU-based approaches, underscoring their adeptness at capturing the contextual nuances essential for discerning fraudulent reviews.

Previous studies utilizing the same dataset of Amazon reviews have reported lower classification performance. For instance, Hajek et al. (Hajek et al., 2020) achieved an accuracy of 82.8% using a Deep Feed Forward Neural Network (DFFNN) method, incorporating features such as BOW n-grams, skip-gram word embeddings, and lexicon-based emotion indicators. Additionally, Birim et al. (Birim et al., 2022) reached an accuracy of 81.58% by using Random Forest classifier and employing features like review length, topic distribution, and purchase validity.

Table 5.1 The Performance of Fake Review Detection Model with Different Embeddings

Embedding	**With RSI**			**Without RSI**		
	Accuracy (%)	AUC (%)	F1 (%)	Accuracy (%)	AUC (%)	F1 (%)
Doc2Vec (C2)	63.06	68.28	61.16	62.1	62.11	63.13
Word2Vec (C1)	61.81	66.55	61.31	61.4	66.77	65.29
GloVe (C1)	**90.21**	**92.58**	**90.07**	87.52	91.65	87.88

Table 5.2 Comparison with Baseline Models

Model	Accuracy (%)	AUC (%)	F1 (%)
BiGRU	63.93	63.91	65.47
CNN	63.64	63.67	60.36
RNN	62.14	63.67	61.9
LSTM	64.09	64.06	65.29
BiLSTM	64.3	64.24	64.95
BiGRU_RSI_GloVe (this study)	**90.21**	**92.58**	**90.07**

Chapter 6 Discussion

6.1 Summary of the Results

This study has successfully demonstrated the effectiveness of incorporating rating-sentiment inconsistency (RSI) into deep learning models for fake review detection. The integration of RSI with pre-trained word embeddings and document embeddings resulted in a notable increase in classification accuracy compared to baseline models. Specifically, the combination of RSI with GloVe embeddings achieved a 3.07% improvement in accuracy, underscoring the potential of RSI as a discriminative feature in identifying deceptive reviews. The results confirm that textual inconsistencies, particularly those involving rating and sentiment, are significant indicators of review authenticity.

6.2 Managerial Implications

The findings of this study have significant implications for e-commerce platforms, consumer review websites, and businesses that rely on online reviews for reputation management and customer engagement. The integration of RSI features into existing review management systems can enhance the ability to detect and filter out fraudulent reviews, thereby maintaining the credibility and

trustworthiness of online content. For e-commerce platforms like Amazon, implementing such models can prevent the manipulation of product ratings, which often leads to distorted consumer perceptions and purchasing decisions.

Additionally, companies can leverage these insights to build more robust monitoring systems that proactively identify potential sources of review fraud, such as bot-generated reviews or coordinated review campaigns. This can help businesses mitigate the adverse effects of fake reviews on their brand reputation and sales performance. Beyond fraud detection, the insights from this research can inform the development of transparency policies and consumer education programs, helping users to better understand and identify fake reviews on their own.

Moreover, the approach of combining textual and inconsistency features can be extended to other domains, such as social media monitoring and customer feedback analysis, where detecting deceitful or biased content is crucial. By investing in advanced fake review detection technologies, businesses can not only protect their online reputation but also ensure a fair competitive environment, which is vital for sustainable market growth.

6.3 Theoretical Implications

The theoretical foundations of cognitive dissonance theory and truth-default theory guided our approach to understanding and applying RSI in detecting fake reviews. Cognitive dissonance theory suggests that users who provide inconsistent information in their ratings and reviews may experience psychological discomfort, leading them to adopt deceptive strategies to maintain a coherent self-image. This behavior aligns with the nature of fake reviews, where positive ratings often conflict with the underlying sentiment, thus creating detectable inconsistencies. Similarly, truth-default theory posits that truthful statements tend to be internally consistent. Therefore, the presence of rating-sentiment inconsistency (as identified through RSI) serves as a potential cue of deception, reinforcing its relevance as an indicator of inauthenticity. By empirically validating RSI's effectiveness, our study bridges these theoretical frameworks with practical application, demonstrating how psychological insights into deceptive behaviors can inform NLP-driven fraud detection.

6.4 Limitations

While the proposed models have demonstrated effectiveness in detecting fake reviews, certain limitations exist:

Feature Limitations: The reliance solely on textual features and RSI may not fully capture the nuanced behavior of sophisticated review fraudsters. Behavioral features, such as user activity patterns, review posting frequency, and network-based analysis, could provide additional insights into identifying deceptive behavior. Moreover, incorporating other textual information, such as syntactic and semantic patterns, contextual word embeddings, and topic modeling, could further enhance the model's capability to detect fake reviews.

Dataset Generalizability: Amazon's filtering system for fake reviews relies on a combination of machine learning algorithms, human reviewers, and customer feedback. This system flags reviews that appear suspicious, based on patterns in review behavior, language, and user activity. Machine learning models identify potential spam by analyzing factors like review timing, frequency, and linguistic cues, while human reviewers and customer reports further validate flagged reviews. However, while Amazon's system is fairly sophisticated, it's not perfect. Reviews flagged as "fake" by this system may still contain some inaccuracies, and authentic reviews may sometimes be incorrectly classified as fake due to complex or unconventional patterns. The other problem with the use of the Amazon dataset, which primarily focuses on product reviews, is that it limits the generalizability of the results to other types of reviews, such as service-based or experience-based feedback. For example, reviews of restaurants, hotels, or freelance services often include subjective opinions and contextual nuances that differ significantly from product reviews. These differences can affect the model's performance and its ability to generalize across various domains. Additionally, the dataset may not reflect the diversity of reviews found on smaller platforms or in non-English languages, where cultural and linguistic variations can play a significant role in the structure and sentiment of reviews.

Challenges with Generative AI: Recent advancements in generative AI, such as GPT-based models, have made it increasingly easy to create highly sophisticated fake reviews that closely mimic genuine user content. These models can produce text with coherent narratives, appropriate emotional tones, and contextually relevant details, making it difficult for traditional detection methods to differentiate between real and fake reviews. This poses a significant challenge for the current research, as the proposed models may not be equipped to handle such advanced, nuanced fake reviews effectively.

Future work could explore integrating adversarial training techniques or leveraging AI-based content verification to tackle this emerging issue.

6.5 Merits of Amazon Dataset

The Amazon dataset provides a large, scalable collection of reviews across diverse products, enhancing model training and generalization. Its labeling process, which combines machine learning, human review, and customer reporting, offers greater reliability compared to manual or crowdsourced methods. As one of the largest e-commerce platforms, Amazon's dataset is practical and relevant, directly applicable to real-world scenarios. The labeling approach, although not flawless, captures realistic fake reviews by mimicking genuine consumer behavior more effectively than generated datasets. However, Amazon's filters are not entirely accurate; some genuine reviews may be flagged as fake and vice versa, introducing noise that can impact model training. The "black box" nature of Amazon's filtering process lacks transparency, posing challenges for researchers who require clear data labeling criteria. Additionally, Amazon's system may carry inherent biases, potentially overlooking certain sophisticated fake reviews, which could limit model effectiveness across varying review types. The dataset's focus on product reviews may restrict the model's applicability to other domains, like service reviews in hospitality or healthcare. Overall, the Amazon dataset balances scalability, reliability, and real-world relevance, though its limitations should be acknowledged for accurate result interpretation.

6.6 Unreliable and fake reviews

Fake reviews are typically considered unreliable because they are intended to mislead rather than provide genuine feedback. They distort consumer perception and can unfairly manipulate product popularity and consumer trust. While a fake review can be spotted through inconsistencies in rating and sentiment, as outlined in this study, the advent of Generative AI presents a new challenge. Advanced AI models, like GPT series, can generate reviews that mimic genuine user opinions with high consistency in both sentiment expression and rating alignment. This not only makes it more challenging to detect such reviews through traditional inconsistency checks but also raises significant concerns about the integrity of user-generated content across platforms. Generative AI can produce high volumes of believable fake content that can sway customer decisions, inflate product ratings, or damage reputations without a trace of inconsistency typically associated with manual fake reviews. As we move forward, it will be essential to develop more sophisticated AI-driven techniques, possibly

employing adversarial AI strategies and deeper linguistic analysis, to detect and counteract these AI-generated fake reviews. This ongoing battle between review authenticity and technological advancements underscores the need for continuous research and adaptation in review verification methodologies.

Chapter 7 Conclusions and future work

Our study presents a comprehensive investigation into the utilization of Rating-Sentiment Inconsistency (RSI) for enhancing fake review detection models. Through meticulous experimentation and analysis, we have delineated two innovative frameworks that integrate this inconsistency with advanced word and document embedding techniques. This research contributes to the field of e-commerce fraud detection by advancing the application of natural language processing and deep learning methodologies. Our proposed frameworks, hinged on the exploitation of RSI, have proven their merit in improving the detection of fake content. The integration of this nuanced textual feature has been empirically validated, underscoring its potential in bolstering the integrity of online review systems.

The enhancements in model performance—substantiated by our rigorous experimental results—affirm the significance of content-based feature integration and the practical application of inconsistency metrics in the realm of fake review detection. The innovative combination of pre-trained embeddings with RSI paves the way for new avenues in NLP applications within commercial settings, providing a blueprint for future research in enhancing online trust and market dynamics. By bridging the gap between theoretical frameworks and real-world applicability, our study stands as a testament to the transformative power of incorporating domain-specific features into NLP tasks. As e-commerce continues to evolve, our findings offer pivotal insights into maintaining the veracity and reliability of user-generated content, ensuring a trustworthy digital marketplace for consumers and businesses alike.

The results show a significant performance difference between models utilizing GloVe embeddings with RSI and those without, as well as compared to other embeddings like Doc2Vec and Word2Vec. This indicates the impact of the choice of word embeddings on model performance. Future studies

could explore the reasons behind this variability, examining how different embeddings capture semantic and syntactic relationships in the text and their suitability for specific classification tasks.

Building on the foundations of this study, future research could significantly expand the scope and depth of understanding in the detection of fake reviews. An important avenue for exploration involves assessing the effectiveness and adaptability of the proposed RSI-based models across a variety of e-commerce platforms and content types. By extending analyses to include datasets from diverse services such as TripAdvisor, Yelp, and Google Reviews, researchers can uncover insights into the models' robustness and fine-tune them for broader application. To improve trust in automated moderation systems, future studies may also investigate explainable AI. By making the decision-making processes of fake review detection models more transparent and interpretable, researchers can help simplify AI operations for users, improving greater confidence in these technologies.

Moreover, conducting longitudinal studies to track the evolution of fake review strategies over time could provide invaluable insights. Such research would offer a window into how fraudulent tactics adapt in response to advances in detection technologies, revealing the dynamic nature of online deception.

Lastly, the emergence of deep fake reviews generated by AI models such as GPT-3 offers a new approach in the spam detection area. Developing detection methods capable of identifying content produced by these advanced technologies will be important. The detection of AI-generated fake content is an emerging and rapidly evolving field. Preliminary research has focused on detecting AI-generated scientific abstracts, employing techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Named Entity Recognition (NER), and Word2Vec for text representation and classification (Sadasivan et al., 2024; Theocharopoulos et al., 2023). These studies indicate that while detection is feasible, the high quality of AI-generated content poses significant challenges. Furthermore, the increasing sophistication of AI language models, such as GPT-3, has made it difficult to distinguish between human-generated and AI-generated texts reliably. As these models continue to advance, the outputs they produce become more similar to human language, complicating detection efforts. Future research should focus on developing more robust and adaptive detection mechanisms. This includes improving existing techniques and exploring new methods that can keep pace with advancements in AI technology. Additionally, interdisciplinary approaches combining AI,

data science, and domain-specific knowledge will be crucial in enhancing the accuracy and reliability of detection systems.

References

Abdulqader, M., Namoun, A., & Alsaawy, Y. (2022). Fake Online Reviews: A Unified Detection Model Using Deception Theories. *IEEE Access*, *10*, 128622–128655. https://doi.org/10.1109/ACCESS.2022.3227631

Abri, F., Gutierrez, L. F., Namin, A. S., Jones, K. S., & Sears, D. R. W. (2020). Linguistic Features for Detecting Fake Reviews. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 352–359. https://doi.org/10.1109/ICMLA51294.2020.00063

Alghamdi, J., Lin, Y., & Luo, S. (2022). Towards Fake News Detection on Social Media. *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, 148–153. https://doi.org/10.1109/ICMLA55696.2022.00028

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Amazon reviews dataset. (2018). [Dataset]. github.com/aayush210789/Deception-Detection-on-Amazon-reviews-dataset

Badri, N., Kboubi, F., & Chaibi, A. H. (2022). Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection. *Procedia Computer Science*, *207*, 769–778. https://doi.org/10.1016/j.procs.2022.09.132

Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, *56*(4), 1234–1244. https://doi.org/10.1016/j.ipm.2019.03.002

Bathla, G., Singh, P., Singh, R. K., Cambria, E., & Tiwari, R. (2022). Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications*, *34*(22), 20213–20229. https://doi.org/10.1007/s00521-022-07531-8

Bhopale, J., Bhise, R., Mane, A., & Talele, K. (2021). A Review-and-Reviewer based approach for Fake Review Detection. *2021 Fourth International Conference on Electrical, Computer and*

Communication Technologies (ICECCT), 1–6. https://doi.org/10.1109/ICECCT52121.2021.9616697

Birim, Ş. Ö., Kazancoglu, I., Kumar Mangla, S., Kahraman, A., Kumar, S., & Kazancoglu, Y. (2022). Detecting fake reviews through topic modelling. *Journal of Business Research*, *149*, 884–900. https://doi.org/10.1016/j.jbusres.2022.05.081

Budhi, G. S., & Chiong, R. (2022). A Multi-type Classifier Ensemble for Detecting Fake Reviews Through Textual-based Feature Extraction. *ACM Transactions on Internet Technology*, 3568676. https://doi.org/10.1145/3568676

Chollet, F. (2018). *Deep learning with Python*. Manning Publications Co.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Doshi, J., Parmar, K., Sanghavi, R., & Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. *Computers & Security*, *133*, 103378. https://doi.org/10.1016/j.cose.2023.103378

Du, J., Rong, J., Wang, H., & Zhang, Y. (2019). Helpfulness Prediction for Online Reviews with Explicit Content-Rating Interaction. In R. Cheng, N. Mamoulis, Y. Sun, & X. Huang (Eds.), *Web Information Systems Engineering – WISE 2019* (Vol. 11881, pp. 795–809). Springer International Publishing. https://doi.org/10.1007/978-3-030-34223-4_50

Elmurngi, E., & Gherbi, A. (2017). Detecting Fake Reviews through Sentiment Analysis Using Machine Learning Techniques. *DATA ANALYTICS*.

Eslami, S. P., & Ghasemaghaei, M. (2018). Effects of online review positiveness and review score inconsistency on sales: A comparison by product involvement. *Journal of Retailing and Consumer Services*, *45*, 74–80. https://doi.org/10.1016/j.jretconser.2018.08.003

Fake online reviews research. (2023). *Department for Business and Trade*.

Five-star fake out. (2022). https://www.canada.ca/en/competition-bureau/news/2022/03/five-star-fake-out.html

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Fraud and scams. (2022). https://competition-bureau.canada.ca/fraud-and-scams

Glazkova, A., & Glazkov, M. (2022). *Detecting Generated Scientific Papers using an Ensemble of Transformer Models* (arXiv:2209.08283). arXiv. http://arxiv.org/abs/2209.08283

GloVe embeddings (Version Apache License, Version 2.0.). (2015). [Dataset]. https://github.com/stanfordnlp/GloVe

Goel, A., Nigam, T., Singh, T., & Agrawal, H. (2021). Classification Of Positive And Negative Fake Online Reviews Using Machine Learning Techniques. *International Journal of Advanced Networking and Applications*, *12*(06), 4746–4749. https://doi.org/10.35444/IJANA.2021.12603

Goldberg, Y., & Levy, O. (2014). *word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method* (arXiv:1402.3722). arXiv. http://arxiv.org/abs/1402.3722

Gutierrez-Espinoza, L., Abri, F., Siami Namin, A., Jones, K. S., & Sears, D. R. W. (2020). Ensemble Learning for Detecting Fake Reviews. *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1320–1325. https://doi.org/10.1109/COMPSAC48688.2020.00-73

Hajek, P., Barushka, A., & Munk, M. (2020). Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, *32*(23), 17259–17274. https://doi.org/10.1007/s00521-020-04757-2

Hajek, P., & Sahut, J.-M. (2022). Mining behavioural and sentiment-dependent linguistic patterns from restaurant reviews for fake review detection. *Technological Forecasting and Social Change*, *177*, 121532. https://doi.org/10.1016/j.techfore.2022.121532

Harmon-Jones, E., & Mills, J. (2019). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology (2nd ed.).* (pp. 3–24). American Psychological Association. https://doi.org/10.1037/0000135-001

Harrison-Walker, L. J., & Jiang, Y. (2023). Suspicion of online product reviews as fake: Cues and consequences. *Journal of Business Research*, *160*, 113780. https://doi.org/10.1016/j.jbusres.2023.113780

He, S., Hollenbeck, B., & Proserpio, D. (2022). The Market for Fake Reviews. *Marketing Science*, *41*(5). https://doi.org/10.1287/mksc.2022.1353

Hmoud Al-Adhaileh, M., & Waselallah Alsaade, F. (2022). Detecting and Analysing Fake Opinions Using Artificial Intelligence Algorithms. *Intelligent Automation & Soft Computing*, *32*(1), 643–655. https://doi.org/10.32604/iasc.2022.021225

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jobin, A., Ienca, M., & Vayena, E. (2019). *Artificial Intelligence: The global landscape of ethics guidelines*.

Kanakaris, N., & Karacapilidis, N. (2023). Predicting prices of Airbnb listings via Graph Neural Networks and Document Embeddings: The case of the island of Santorini. *Procedia Computer Science*, *219*, 705–712. https://doi.org/10.1016/j.procs.2023.01.342

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, *90*, 523–537. https://doi.org/10.1016/j.indmarman.2019.08.003

Kotkov, D., Medlar, A., Satyal, U. R., Maslov, A., Neovius, M., & Glowacka, D. (2022). Rating consistency is consistently underrated: An exploratory analysis of movie-tag rating inconsistency. *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 1355–1364. https://doi.org/10.1145/3477314.3507270

LeCun, Y. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1566–1576. https://doi.org/10.3115/v1/P14-1147

Li, L., Lee, K. Y., Lee, M., & Yang, S.-B. (2020). Unveiling the cloak of deviance: Linguistic cues for psychological processes in fake online reviews. *International Journal of Hospitality Management*, *87*, 102468. https://doi.org/10.1016/j.ijhm.2020.102468

Liashchynskyi, P., & Liashchynskyi, P. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS* (arXiv:1912.06059). arXiv. http://arxiv.org/abs/1912.06059

Lin, T.-Y., Chakraborty, B., & Peng, C.-C. (2021). A Study on Identification of Important Features for Efficient Detection of Fake Reviews. *2021 International Conference on Data Analytics*

for Business and Industry (ICDABI), 429–433. https://doi.org/10.1109/ICDABI53623.2021.9655845

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Liu, Y., Wang, L., Shi, T., & Li, J. (2022). Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Information Systems*, *103*, 101865. https://doi.org/10.1016/j.is.2021.101865

Luvembe, A. M., Li, W., Li, S., Liu, F., & Xu, G. (2023). Dual emotion based fake news detection: A deep attention-weight update approach. *Information Processing & Management*, *60*(4), 103354. https://doi.org/10.1016/j.ipm.2023.103354

Marciano, J. (2021). *Fake online reviews cost $152 billion a year. Here's how e-commerce sites can stop them*. https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/

Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake Reviews Detection: A Survey. *IEEE Access*, *9*, 65771–65802. https://doi.org/10.1109/ACCESS.2021.3075573

Mridha, M. F., Keya, A. J., Hamid, Md. A., Monowar, M. M., & Rahman, Md. S. (2021). A Comprehensive Review on Fake News Detection With Deep Learning. *IEEE Access*, *9*, 156151–156170. https://doi.org/10.1109/ACCESS.2021.3129329

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st International Conference on World Wide Web*, 191–200. https://doi.org/10.1145/2187836.2187863

Nagi Alsubari, S., N. Deshmukh, S., Abdullah Alqarni, A., Alsharif, N., H. H. Aldhyani, T., Waselallah Alsaade, F., & I. Khalaf, O. (2022). Data Analytics for the Identification of Fake Reviews Using Supervised Learning. *Computers, Materials & Continua*, *70*(2), 3189–3204. https://doi.org/10.32604/cmc.2022.019625

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. *Proceedings of the 21st International Conference on World Wide Web*, 201–210. https://doi.org/10.1145/2187836.2187864

Ott, M., Cardie, C., & Hancock, J. T. (n.d.). *Negative Deceptive Opinion Spam*.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Qu, X., Li, X., & Rose, J. R. (2018). *Review Helpfulness Assessment based on Convolutional Neural Network* (arXiv:1808.09016). arXiv. http://arxiv.org/abs/1808.09016

Qu, Y., Zhang, W. E., Yang, J., Wu, L., & Wu, J. (2022). Knowledge-aware document summarization: A survey of knowledge, embedding methods and architectures. *Knowledge-Based Systems*, *257*, 109882. https://doi.org/10.1016/j.knosys.2022.109882

Rao, S., Verma, A. K., & Bhatia, T. (2023). Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data. *Expert Systems with Applications*, *217*, 119594. https://doi.org/10.1016/j.eswa.2023.119594

Refaeli, D., & Hajek, P. (2021). Detecting Fake Online Reviews using Fine-tuned BERT. *2021 5th International Conference on E-Business and Internet*, 76–80. https://doi.org/10.1145/3497701.3497714

Ren, Y., & Ji, D. (2017). Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, *385–386*, 213–224. https://doi.org/10.1016/j.ins.2017.01.015

Rout, J. K., Dash, A. K., & Ray, N. K. (2018). A Framework for Fake Review Detection: Issues and Challenges. *2018 International Conference on Information Technology (ICIT)*, 7–10. https://doi.org/10.1109/ICIT.2018.00014

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/doi.org/10.1038/323533a0

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2024). *Can AI-Generated Text be Reliably Detected?* (arXiv:2303.11156). arXiv. http://arxiv.org/abs/2303.11156

Schuman, A. (2022). *What fake reviews can mean for your business*. https://www.bazaarvoice.com/blog/what-fake-reviews-can-mean-for-your-business/

Schuster, M., & Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Shan, G., Zhou, L., & Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems*, *144*, 113513. https://doi.org/10.1016/j.dss.2021.113513

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306

Singh, D., Memoria, M., & Kumar, R. (2023). Deep Learning Based Model for Fake Review Detection. *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 92–95. https://doi.org/10.1109/InCACCT57535.2023.10141826

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (n.d.). *Release Strategies and the Social Impacts of Language Models*.

Study examines the impact of fake online reviews on sales. (2022). https://phys.org/news/2022-09-impact-fake-online-sales.html

The Deceptive Marketing Practices Digest—Volume 1. (2015). https://competition-bureau.canada.ca/deceptive-marketing-practices-digest-volume-1#s3_0

Theocharopoulos, P. C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S. V., Tasoulis, S. K., & Plagianakos, V. P. (2023). Detection of Fake Generated Scientific Abstracts. *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, 33–39. https://doi.org/10.1109/BigDataService58306.2023.00011

Thompson, R. C., Joseph, S., & Adeliyi, T. T. (2022). A Systematic Literature Review and Meta-Analysis of Studies on Online Fake News Detection. *Information*, *13*(11), 527. https://doi.org/10.3390/info13110527

Wang, S., Lin, Y., & Zhu, G. (2023). Online reviews and high-involvement product sales: Evidence from offline sales in the Chinese automobile industry. *Electronic Commerce Research and Applications*, *57*, 101231. https://doi.org/10.1016/j.elerap.2022.101231

Wang, Y., Ngai, E. W. T., & Li, K. (2023). The effect of review content richness on product review helpfulness: The moderating role of rating inconsistency. *Electronic Commerce Research and Applications*, *61*, 101290. https://doi.org/10.1016/j.elerap.2023.101290

Wang, Y., Zamudio, C., & Jewell, R. D. (2023). The more they know: Using transparent online communication to combat fake online reviews. *Business Horizons*, *66*(6), 753–764. https://doi.org/10.1016/j.bushor.2023.03.004

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text.

International Journal for Educational Integrity, *19*(1), 26. https://doi.org/10.1007/s40979-023-00146-z

Yao, J., Zheng, Y., & Jiang, H. (2021). An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization. *IEEE Access*, *9*, 16914–16927. https://doi.org/10.1109/ACCESS.2021.3051174

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing* (arXiv:1702.01923). arXiv. http://arxiv.org/abs/1702.01923

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2020). *Defending Against Neural Fake News* (arXiv:1905.12616). arXiv. http://arxiv.org/abs/1905.12616

Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews. *Journal of Management Information Systems*, *33*(2), 456–481. https://doi.org/10.1080/07421222.2016.1205907