CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Since the advent of the internet and social networks, Sentiment Analysis has evolved into the most important area of research in Natural Language Processing. Sentiment Analysis not only identifies user emotions and the direction of public opinion but also plays an important role in application areas like product recommendation, market promotion, and financial prediction. In the last several years, the availability of large-scale User-Generated Content (UGC) from diverse sources such as Yelp, Amazon, IMDB, and Twitter has greatly enriched the data resource for sentiment analysis. The expansion has also brought about new challenges in terms of requirements for finer-grained labeling, increased data variety, as well as more sophisticated business needs. Thus, it is necessary to revisit comprehensively mainstream datasets, modeling methods, explainability techniques, and misclassification analysis theories to shed light on the development trajectory of the field, comprehend state-of-the-art challenges, and outline future research directions.

This chapter will systematically elaborate on the key issues like multi-source sentiment analysis datasets, traditional and deep learning approaches, explainability analysis, and multi-class misclassification measure. Through summarizing and contrasting the existing research attainments, we aim to explore the common issues and technical challenges in the task of multi-class sentiment modeling. The chapter will ultimately formulate key unsolved research gaps with a vision of laying a theoretical foundation and practical groundwork for offering follow-up research contents and developing novel approaches.

2.2 Application Background and Motivation

In recent years, the rapid advancement of the internet and Web 2.0 technologies has driven a full-scale explosion of User-Generated Content (UGC). Users actively publish reviews about products, services, and experiences on social platforms, review websites, and forums. This UGC has become an indispensable resource for information dissemination, consumer decision-making, and business management in modern society (Almansour et al., 2022). Statistics show that over 75% of consumers actively check online reviews before making purchasing decisions, and most users consider these reviews more valuable than advice from friends and family (Almansour et al., 2022).

Among the many UGC platforms, Yelp stands out for its focus on local life services, dining, and entertainment, accumulating a vast amount of high-quality review data. What makes Yelp unique is that, in addition to user-written text reviews, it systematically integrates multi-dimensional structured information such as user ratings (star ratings), timestamps, geographical locations, business categories, price ranges, and user activity. This provides an ideal scenario and rich samples for multi-source data fusion and large-scale sentiment analysis research (Alamoudi & Alghamdi, 2021; Lak & Turetken, 2014).

However, the explosive growth of UGC reviews has also brought unprecedented challenges. Firstly, information overload is increasingly prominent, making it difficult for users and businesses to efficiently filter truly valuable review content (Almansour et2 al., 2022). Secondly, UGC content is highly subjective, with significant variations in expression, length, and focus; some reviews are detailed and useful, while others are off-topic or lack practical reference value (Alamoudi & Alghamdi, 2021). More complexly, the actual emotional polarity in user ratings and their text reviews is not always consistent. This "Text-Rating Review Discrepancy" (TRRD) problem has been confirmed by multiple studies, and simply using ratings as sentiment labels can mislead subsequent model training and prediction (Almansour et al., 2022; Lak & Turetken, 2014).

Facing these challenges, there's a growing urgent demand from both society and academia for "fine-grained, multi-faceted sentiment analysis." Traditional sentiment analysis methods primarily focus on overall polarity judgment. However, in real life and business decisions, users are often more concerned about specific evaluations of reviews across different aspects (e.g., taste, ambiance, service, and value for money in the catering industry). For instance, a single review might contain multi-dimensional opinions such as "the dishes were delicious but the service was poor," where a single polarity label cannot accurately reflect its complex emotional structure (Fan & Zhang, 2024; Alamoudi & Alghamdi, 2021).

Therefore, sentiment analysis of Yelp review data not only helps consumers make more informed choices but also provides crucial data support for businesses in precision marketing, reputation management, and product improvement. Simultaneously, advancing cutting-edge topics such as multi-source information fusion, fine-grained sentiment recognition, and text-rating consistency modeling has become a mainstream research direction and a hot topic in industry practice within the field of sentiment analysis (Fan & Zhang, 2024; Almansour et al., 2022).

2.3 Overview and Comparison of Main Sentiment Analysis Datasets

For the diverse tasks of sentiment analysis, academia and industry have accumulated a variety of high-quality public datasets, supporting the continuous evolution of models and methods.

2.3.1 Yelp Dataset

The Yelp dataset is one of the most widely used real-world review big data in the current research fields of sentiment analysis and recommendation systems. This dataset was first launched by Yelp and has been continuously updated and upgraded. It currently contains millions of user reviews, star ratings, user information, and business attributes from industries such as catering, leisure, and entertainment worldwide. Yelp data is not only highly structured but also includes rich

metadata such as free text, timestamps, geographical information, price ranges, and tags, making it very suitable for multi-source data modeling, fine-grained sentiment analysis, and explainability research (Alamoudi & Alghamdi, 2021). Yelp's labeling system mainly uses 1–5 star ratings, supporting multi-class classification and regression analysis. It has high representativeness and application value in cutting-edge research such as multi-aspect polarity modeling, star-text inconsistency, and UGC multimodal mining (Fan & Zhang, 2024).

The Yelp Academic Dataset can be downloaded for free after registration on its official website (https://www.yelp.com/dataset) for academic or non-commercial research use only. The official website will update the data version from time to time, which is convenient for researchers to reproduce and compare methods (Alamoudi & Alghamdi, 2021).

2.3.2 Amazon Review Dataset

The Amazon Review dataset is another large-scale, broadly diversified, general-purpose sentiment analysis dataset. This dataset aggregates tens of millions of user reviews and rating information from various industries on the Amazon platform (such as books, digital products, apparel, home goods, etc.). It features detailed structured data including product IDs, categories, text, ratings (1–5 stars), timestamps, and user behavior. Compared to Yelp, the Amazon Review dataset exhibits stronger cross-category and spatio-temporal characteristics, supporting more complex applications like domain transfer, product recommendations, and user behavior modeling (Katić & Milićević, 2018). Its primary challenges include uneven review distribution, a higher prevalence of fake reviews, and numerous outliers. Nevertheless, its sheer scale and diversity offer significant value for academic and engineering research (He & McAuley, 2016).

The Amazon review dataset can be obtained through various channels, such as Kaggle, AWS Public Datasets, or researchers' homepages. The most classic "Amazon Product Data" is continuously maintained by the McAuley team (https://nijianmo.github.io/amazon/index.html),

offering convenient downloads, clear categorization, and applicability to various specific tasks (He & McAuley, 2016).

2.3.3 IMDB Movie Review Dataset

The IMDB Movie Review dataset is one of the earliest and most standard benchmarks in the field of text sentiment analysis. First compiled and released by Maas et al., this dataset primarily consists of free text movie reviews with positive and negative labels, making it suitable for binary sentiment classification and text feature extraction experiments. The IMDB dataset has a simple structure, typically features longer text lengths, and generally does not include additional user or product metadata. Its main advantages are clean data and accurate labeling, which makes it ideal for baseline testing of new algorithms and preliminary sentiment analysis research (Maas et al., 2011). However, compared to Yelp and Amazon datasets, its extensibility is limited for tasks such as multi-class classification, fine-grained analysis, and multi-source fusion.

The IMDB sentiment analysis dataset can be downloaded for free from the Stanford NLP team's official website (https://ai.stanford.edu/~amaas/data/sentiment) required. Its standardized format and clear labels make it highly suitable for academic experiments (Maas et al., 2011).

2.3.4 X (Twitter) Dataset

The X (Twitter) dataset is an important benchmark for social media text analysis and sentiment computing. This type of dataset primarily consists of tweets, which are short, information-dense texts covering various topics such as politics, entertainment, current events, and product feedback. X sentiment analysis primarily focuses on tasks like positive-negative-neutral three-class classification, topic extraction, and network event evolution, with common datasets including Sentiment140 and SemEval. X data is characterized by strong timeliness, large volume, diverse language styles, and rich emojis. However, it also faces challenges such as high text noise,

complex implicit emotional expressions, and frequent use of irony and puns (Wang et al., 2022; Qi & Shabrina, 2023). X data is highly suitable for cutting-edge research in multi-task learning, sentiment trend mining, and online emotion prediction.

Typical X (Twitter) sentiment datasets like Sentiment140 and SemEval can be downloaded for free from project homepages or the ACL data platform. However, the original tweet content needs to be scraped via the X API (requiring a developer account). Due to platform policies and tweet ID validity, data collection has certain technical barriers, but it is highly open and globally applicable (Wang et al., 2022; Qi & Shabrina, 2023).

2.3.5 Comparative Analysis of Four Major Datasets

In order to more intuitively demonstrate the similarities and differences in structure, label types, and applicable scenarios among the four major sentiment analysis datasets, their core features are summarized in the table below.

Table 2.3.5 Overview of Main Public Sentiment Analysis Datasets

Dateset	Data Size	Domain	Label Type	Metadata	Typical Text Length	Representative Applications	Main Challenges	References
Yelp	Millions	Local Life/Dining	1–5 stars	Rich	Medium	Multiclass, fine-grained sentiment, explainability	Text-rating inconsistency, high subjectivity	Alamoudi & Alghamdi (2021); Fan & Zhang (2024)
Amazon	Tens of millions	All product categories	1–5 stars	Rich	Medium–Long	Recommendation, domain adaptation, multi-source analysis	Uneven distribution, fake/extreme reviews	Katić & Milićević (2018); He &

								McAuley (2016)
IMDB	Tens to hundreds of thousands	Movies	Binary	Simple	Long	Binary classification, feature extraction, baseline testing	Lack of diversity, limited scalability	Maas et al. (2011)
X (Twitter)	Millions	Social domain	Three/multiclass	General	Short	Opinion mining, event tracking, NLP benchmarks	High noise, sarcasm, complex expressions	Wang et al. (2022); Qi & Shabrina (2023)

As summarized above, mainstream public datasets like Yelp, Amazon, IMDB, and X (Twitter) each have their own strengths in terms of data scale, domain focus, labeling systems, and metadata richness, providing a solid foundation for sentiment analysis research. Selecting a dataset with matching characteristics for a specific research question is crucial for model performance, generalization ability, and the effectiveness of the final application (Katić & Milićević, 2018; He & McAuley, 2016; Wang et al., 2022).

Beyond these standard datasets, dynamically collecting raw reviews or news data for specific scenarios through techniques like web scraping has also become an important trend. This approach helps improve the timeliness and task specificity of the data, and it can enhance researchers' comprehensive abilities in data engineering and practical business scenarios (Kaur, 2022).

Therefore, future research will comprehensively consider the applicability of existing public datasets and the feasibility of self-collected data, flexibly designing data acquisition and preprocessing workflows.

2.4 Overview and Comparative Analysis of Sentiment Analysis Methods

As sentiment analysis tasks continue to evolve, researchers have proposed various modeling approaches. Based on different technical routes and theoretical foundations, mainstream methods can be broadly categorized into three types: traditional machine learning methods, classic methods specifically designed for sentiment analysis, and deep learning and pre-trained model methods. A systematic review of each will be provided below.

2.4.1 Traditional Machine Learning Approaches

1. Naive Bayes (NB)

Naive Bayes (NB) is a probabilistic classifier widely used for text sentiment classification due to its simplicity, efficiency, and effectiveness with high-dimensional, sparse features. The core assumption of NB is that features (such as words or n-grams in a document) are conditionally independent given the class. In sentiment analysis, NB calculates the posterior probability of each sentiment class given the observed features, assigning the class with the highest probability as the final label (Arya et al., 2022; Fransisca et al., 2021; Ghatora et al., 2024; Ramasamy et al., 2023; Das et al., 2023).

The most representative classification formula for Naive Bayes in text sentiment analysis is as follows:

$$\hat{y} = \arg\max_{y \in Y} P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

where $\hat{y}$ is the predicted sentiment label, Y is the set of possible sentiment classes (e.g., positive, negative, neutral), x_i represents the i th feature (word) in the text, $P(y)$ is the prior

probability of class y, and $P(x_i \mid y)$ is the conditional probability of observing feature x_i given class y (Ghatora et al., 2024; Fransisca et al., 2021).

Naive Bayes has also been found to work well with short-text and moderately sized datasets for languages and platforms (Arya et al., 2022; Das et al., 2023). Its principal limitation, though, is the strong independence assumption, which can be untrue for natural language; thus, its performance deteriorates in the presence of highly correlated features or subtle/complicated sentiment expressions (Fransisca et al., 2021; Das et al., 2023; Ghatora et al., 2024).

2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a discriminative classifier that seeks to determine the best separating hyperplane with maximum margin between classes in a high-dimensional feature space. SVM is very efficient for text classification problems like sentiment analysis, particularly in the case of sparse and high-dimensional data (Han et al., 2020; Das et al., 2023; Ghatora et al., 2024; Singh et al., 2022; Benarafa et al., 2024).

The SVM classification decision function used in sentiment analysis is given by:

$$f(z) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, z) + b\right)$$

where z is the feature vector of the text to be classified, x_i are support vectors from the training set, y_i are their class labels, α_i are the learned weights, $K(x_i, z)$ is the kernel function measuring similarity, b is the bias term, and sgn denotes the sign function which assigns the final class (Benarafa et al., 2024; Han et al., 2020).

SVM has demonstrated superior performance compared to Naive Bayes and other traditional models, particularly for binary and multi-class sentiment classification tasks with short

texts, such as product reviews and tweets (Das et al., 2023; Ramasamy et al., 2023; Ghatora et al., 2024). Advanced versions using kernel tricks (such as RBF, polynomial, or semantic kernels) further enhance its ability to capture non-linear relationships and implicit sentiment aspects (Benarafa et al., 2024; Han et al., 2020). However, SVM can be computationally intensive for very large datasets and requires careful tuning of hyperparameters and kernel selection (Ghatora et al., 2024; Ramasamy et al., 2023).

3. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, instance-based learning algorithm commonly used as a baseline for sentiment analysis tasks. Its core principle is to classify an unseen text sample by the majority sentiment label among its k closest neighbors in the feature space, based on a similarity or distance measure (Abo et al., 2021; Saudy et al., 2022).

KNN requires no explicit training phase; instead, it stores all training samples and classifies new samples on demand. Typical workflows in sentiment analysis include text vectorization (e.g., TF-IDF), normalization, and the application of KNN for multi-class sentiment prediction. KNN is valued for its simplicity, interpretability, and ease of implementation, especially for small to medium-sized datasets where category boundaries are distinct (Saudy et al., 2022).

However, KNN suffers from several limitations: its computational cost increases rapidly with dataset size, it is sensitive to the curse of dimensionality and noise, and its performance strongly depends on the choice of distance metric and value of k (Abo et al., 2021). In empirical studies on Arabic sentiment analysis and mobile app review classification, KNN often achieved lower accuracy and F1-scores compared to advanced models such as Random Forest, Logistic Regression, and deep learning approaches, but it remains a valuable baseline for algorithm comparison and hybrid model integration (Saudy et al., 2022; Abo et al., 2021).

4. Logistic Regression (LR)

Logistic Regression (LR) is a popular linear classifier used for sentiment analysis, particularly effective for large-scale and high-dimensional text data. For multiclass sentiment classification tasks (e.g., Yelp 1–5 stars), LR is typically extended to multinomial logistic regression using the softmax function, modeling the probability that a text belongs to each of the K sentiment classes:

$$P(\boldsymbol{y}=\boldsymbol{k}|\boldsymbol{X}) = \frac{\exp(\beta_{k,0} + \sum_{i=1}^{n} \beta_{k,i}\boldsymbol{x}_i)}{\sum_{j=1}^{K} \exp(\beta_{j,0} + \sum_{i=1}^{n} \beta_{j,i}\boldsymbol{x}_i)}$$

where $P(\boldsymbol{y}=\boldsymbol{k}|\boldsymbol{X})$ is the probability of assigning class $\boldsymbol{k}$ to text sample $\boldsymbol{X}$, $\beta_{k,0}$ and $\beta_{k,i}$ are the bias and feature weights for class k, and $\boldsymbol{x}_i$ are the text features (Wenping Wang et al., 2023; Singh & Jaiswal, 2023).

Empirical findings indicate that multinomial logistic regression, particularly when coupled with vectorization methods like TF-IDF, is competitive in multiclass sentiment analysis tasks like Yelp and Twitter reviews. It is as good as, if not superior to, more sophisticated models with the added advantage of fast inference speed and interpretability (Wenping Wang et al., 2023; Padhy et al., 2024). Yet, its linear modeling capacity could restrict its effectiveness in capturing sophisticated patterns of sentiment when feature interactions and context are important (Singh & Jaiswal, 2023).

5. Decision Trees (DT)

Decision Trees (DTs) are popular non-parametric supervised multiclass sentiment analysis models appreciated for their interpretability and ability to deal with categorical and numerical

features. For sentiment classification, DTs learn a hierarchical tree model by recursively dividing the feature space—e.g., bag-of-words, TF-IDF, or word embedding vectors—according to feature values maximising a split criterion.

At each internal node, the algorithm chooses a feature and a threshold that give the best discrimination among sentiment classes based on measures like information gain or the Gini index (Dandash & Asadpour, 2023; Jain et al., 2023).

An example of a split criterion representative for the CART decision tree algorithm is the Gini index:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

where $Gini(D)$ is the impurity of the node containing dataset D, K is the number of sentiment classes, and p_k is the proportion of samples belonging to class k at that node. The algorithm selects the split that results in child nodes with minimal weighted Gini impurity.

Decision Trees are attractive for sentiment analysis due to their straightforward structure and ability to visualize the decision-making process—helping reveal which words or features most strongly affect the sentiment classification. Empirical results demonstrate that DTs, when combined with appropriate feature representations, can perform competitively on multiclass sentiment analysis tasks. For example, Dandash & Asadpour (2023) report that DTs, applied to Arabic social media sentiment classification with bag-of-words and TF-IDF features, achieved accuracies ranging from 22% to 38% across various multiclass settings. Jain et al. (2023) highlight DTs as a key component in multimedia sentiment analysis pipelines, and in broader benchmark comparisons, decision trees often serve as interpretable baselines or as base learners in ensemble methods (such as Random Forests and Gradient Boosted Trees) (Dandash & Asadpour, 2023; Jain et al., 2023).

However, DTs are prone to overfitting in high-dimensional and sparse feature spaces, especially common in text data, and their performance may be surpassed by more robust algorithms like SVM, logistic regression, or deep learning models. Nevertheless, their transparency and ability to natively handle multiclass tasks make them valuable tools for both standalone and ensemble sentiment classification systems.

6. Random Forest (RF)

Random Forest (RF) is an ensemble approach whereby a huge number of decision trees are trained on the train data at train time and their predictions aggregated together for impro ving classification accuracy and generalization. RF grows each tree on a bootstrapped version of the train data and uses a random subset of features at each split, making the ensemble stronger and less vulnerable to overfit compared to single decision trees (Dandash & Asadpour, 2023; Jain et al., 2023). The final sentiment class for a given instance is determined based on the voting decision of all of the forest's trees:

$$\hat{y} = \text{majority\_vote}\{h_t(\mathbf{x}) \mid t = 1, \ldots, T\}$$

where $\hat{y}$ is the sentiment prediction for input vector $\mathbf{x}$, $h_t(\mathbf{x})$ is the t-th tree prediction and T is the total number of trees.

Random Forest has been found suited well for sentiment classification on text-based high-dimensional bag-of-words, TF-IDF, or n-grams. Empirical work demonstrates RF working reliably on multi-class sentiment analysis. Dandash & Asadpour (2023), for instance, established the performance of RF superior compared to single decision trees, KNN, and a number of linear classifiers on Arabic sentiment corpora from Twitter and Facebook containing as high as 40.54% accuracy from character-level TF-IDF features. Jain et al. (2023) also cite RF as a competitive and popular baseline for sentiment analysis and multimedia pipelines and report competitive performance on a variety of benchmark datasets.

Despite these strengths, Random Forest can be computationally demanding as the number of trees and features increases, and while it supports feature importance analysis, its decision process is less transparent than that of a single tree. Nonetheless, RF remains one of the most popular and effective methods for multiclass sentiment analysis, valued for its predictive power and reliability in practical applications (Dandash & Asadpour, 2023; Jain et al., 2023).

7. Model Comparison

Traditional machine learning models commonly used in sentiment analysis each have their own characteristics. The table below compares the main advantages, limitations, typical application scenarios, and relevant references of these models.

Table 2.4.1.7 Comparison of Traditional Machine Learning Models for Sentiment Analysis

Model	Advantages	Limitations	Typical Use Case	References
Naive Bayes	Fast, interpretable, handles sparse data well	Independence assumption, lower accuracy on complex data	Short/simple texts, baseline	Arya et al., 2022; Ghatora et al., 2024
Support Vector Machine	High accuracy, strong generalization	Computationally intensive, needs tuning	Binary and multiclass text classification	Han et al., 2020; Benarafa et al., 2024
K-Nearest Neighbors	Simple, no explicit training	Not scalable, sensitive to noise	Small datasets	Abo et al., 2021; Saudy et al., 2022
Logistic Regression	Efficient, interpretable, good with multiclass	Linear, limited with complex non-linear patterns	Large, multiclass datasets	Wenping Wang et al., 2023; Singh & Jaiswal, 2023
Decision Trees	Interpretable, supports multiclass	Overfitting, less robust on large/sparse data	Visualization, analysis	Dandash & Asadpour, 2023; Jain et al., 2023
Random Forest	High accuracy, robust, reduces overfitting	Less interpretable, resource-intensive with many trees	Complex, multiclass tasks, feature importance	Dandash & Asadpour, 2023; Jain et al., 2023

As evident from the table, different models vary in terms of accuracy, training efficiency, and applicable scenarios. Naive Bayes and Logistic Regression are suitable for high-dimensional sparse text due to their computational efficiency, but they have limitations in modeling complex relationships. SVM and Random Forest excel in classification performance but demand higher computational resources. In practical applications, it's crucial to select models judiciously, considering data characteristics and project requirements to achieve a balance between performance and efficiency.

2.4.2 Sentiment-specific Classic Methods

Common dictionary and rule-based methods used in sentiment analysis include SentiWordNet, VADER, and TextBlob. These methods don't rely on large amounts of manually labeled data; instead, they determine the overall sentiment polarity of a text through pre-built sentiment lexicons combined with simple score aggregation or rule-based judgments (Taboada et al., 2011; Qi & Shabrina, 2023).

1. Principles of Mainstream Dictionary-Based Methods

- SentiWordNet assigns positive, negative, and neutral scores to each word or phrase based on WordNet's synsets. It is suitable for general English text but has limitations in handling internet slang and emoticons (Nursal et al., 2025).

- VADER is specifically designed for social media, with its lexicon covering internet slang, emojis, and colloquialisms. It also incorporates rules (e.g., negation, exclamation, lexical emphasis) to adjust sentiment intensity, making it particularly effective for short texts like those found on Twitter and forums (Qi & Shabrina, 2023; Nursal et al., 2025).

- TextBlob primarily scores each word or phrase based on its lexicon, then calculates the average overall polarity. It's suitable for general texts and product reviews (Qi & Shabrina, 2023).

2. Empirical Comparisons and Application Characteristics

A recent study (Nursal et al., 2025) compared the performance of WordNet, SentiWordNet, TextBlob, and VADER on Malaysian high-rise residential forum and Google review data, revealing the following:

- All lexicon-based methods tend to identify more positive sentiment, but VADER performs better in identifying negative and neutral sentiment, achieving the highest overall classification accuracy (78%) and a recall rate of up to 90%.

- SentiWordNet and WordNet have broad coverage but limited adaptability to slang and new words. They are susceptible to social media noise, resulting in slightly lower accuracy than VADER and TextBlob.

- TextBlob's overall performance is moderate, suitable for general English text analysis, but it's somewhat limited when handling informal text and nuanced emotions.

Moreover, common advantages of lexicon-based methods include simple implementation, computational efficiency, ease of interpretation, independence from training samples, suitability for cold-start and low-resource scenarios, or as a baseline for machine learning models (Ghatora et al., 2024; Nursal et al., 2025). However, they also have clear shortcomings: difficulty in recognizing complex expressions such as sarcasm, negation, ambiguity, and spelling errors, and limitations in lexicon coverage can affect adaptability to new domains (Nursal et al., 2025).

2.4.3 Deep Learning and Pre-trained Language Models

2.4.3.1 Basic Deep Learning Methods

In recent years, deep learning techniques have made significant strides in the field of sentiment analysis. Traditional machine learning methods rely on manual feature extraction, whereas deep learning models can automatically learn complex semantic and temporal features from raw text, greatly improving the accuracy of sentiment classification (Wu et al., 2022; Jin, 2023).

1. Convolutional Neural Networks (CNN)

RNNs and their extensions (LSTM, GRU, and BiLSTM) can capture contextual dependency in a sequence of text. The vanishing and exploding gradient problem of traditional RNNs when dealing with long sequences can be overcome by LSTM and GRU through gate mechanism and thus they are better suited for capturing long texts and semantically highly consistent contexts. BiLSTM also employs bidirectional information and thus inherits their ability to capture contextual sentiment (Golubeva & Loukachevitch, 2021; Wu et al., 2022).

2. Recurrent Neural Networks (RNN) and Variants

The researchers have proposed a series of ensemble architecture models for better sentiment analysis performance. As an example, CNN-LSTM combines the advantage of CNN in local feature extraction and the advantage of LSTM in temporal dependency modeling so as to achieve a good sense of sentiment of text (Wu et al., 2022). Similarly, multi-level models like Co-LSTM and Two-Level LSTM enhance the advantage of the model in grasping complex emotional expressions (e.g., inversions and negations) through the addition of sentiment lexicons or polarity reversal (Wu et al., 2022; Jin, 2023). Meanwhile, the integration of LSTM and GRU has also seen fruitful applications in a plethora of areas ranging from cryptocurrency and finance to healthcare (Jin, 2023).

3. Ensemble and Hybrid Models

The researchers have proposed a series of ensemble architecture models for better sentiment analysis performance. As an example, CNN-LSTM combines the advantage of CNN in local feature extraction and the advantage of LSTM in temporal dependency modeling so as to achieve a good sense of sentiment of text (Wu et al., 2022). Similarly, multi-level models like Co-LSTM and Two-Level LSTM enhance the advantage of the model in grasping complex emotional expressions (e.g., inversions and negations) through the addition of sentiment lexicons or polarity reversal (Wu et al., 2022; Jin, 2023). Meanwhile, the integration of LSTM and GRU has also seen fruitful applications in a plethora of areas ranging from cryptocurrency and finance to healthcare (Jin, 2023).

2.4.3.2 Pre-trained Language Models

Over the last few years, sentiment analysis has also seen a big boost through pre-trained Transformer language models. Pre-trained models like BERT can capture more semantic and contextual relationships through unsupervised pretraining on a large corpus of texts and hence significantly enhance the performance of downstream applications like sentiment classification (Devlin et al., 2019).

1. BERT and its Variants

BERT employs a multi-layer bidirectional Transformer encoder capable of simultaneously capturing both left and right context words of a given text. BERT introduced two-stage unsupervised pre-training tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). BERT significantly enhanced the precision and generalization capacity of sentiment classification via its "pre-training + fine-tuning" mechanism (Devlin et al., 2019).

BERT and variants thereof (e.g., RoBERTa, DistilBERT, and BERTweet) also demonstrated high cross-linguistic and cross-domain generalizability. Multilingually trained models like RuBERT and GreekBERT also demonstrated excellent performance in sentiment analysis for Russian and Greek languages (Syrigka et al., 2023; Golubeva & Loukachevitch, 2021).

2. Multi-Model Fusion and Architectural Innovations

With increasing application requirements, researchers begin to combine BERT-type models with traditional deep learning models (e.g., CNN, BiLSTM, ResNeXt, etc.) for enhanced feature extraction capacity and better performance of the model. As a case in point, triple fusion approach RoBERTa-ResNeXt-BiLSTM results in higher customer review sentiment analysis precision compared to standalone models (Farah et al., 2024). Dual-channel deep classifier (DJC) of RoBERTa and BERT also can effectively handle imbalanced data and multi-category sentiment and improve the generalization capacity and stability of the model (Zhang et al., 2024).

3. Comparative Analysis and Applications of Various Pre-trained Models

Recent research indicates BERT and BERT variant models (e.g., RoBERTa, GPT-2, XLNet) performing better than traditional machine learning and shallow deep-learning models on various publicly available datasets such as IMDB, Yelp, Twitter, and clinical medical discussions (Syrigka et al., 2023). Amongst them, RoBERTa excels at long-text and fine-grained polarity responses, BERT would be naturally fit for multi-lingual contexts, and the generative models GPT-2 and XLNet perform well on emotional reasoning and text generation tasks.

2.5 Overview of Explainability Methods in Sentiment Analysis

2.5.1 The Evolution of Explainability Techniques

With deep learning models increasingly used in sentiment analysis, the "black-box" phenomenon has become a prominent concern recently. In other words, the models can make predictions but cannot as clearly explain the decision-making foundations behind their predictions, which hinders their application in high-risk sectors and actual business operations. Consequently, a surge in explainability (Explainability/Interpretability) research has emerged in the AI and NLP fields, driving the birth of various interpretability techniques (Ghasemi & Momtazi,

2023). On one hand, as a task closely related to subjective user experience, sentiment analysis results often influence business decisions, policy evaluations, and even financial trends, making it particularly crucial to "make users trust why the model made a certain sentiment judgment" (Rizinski et al., 2024). On the other hand, legal regulations (such as the EU GDPR) also mandate that some AI systems must possess traceability and explainability for their results, which has further driven the development of relevant methods (Tutek & Šnajder, 2022).

2.5.2 Mainstream Explainability Methods Categorization

1. Local Interpretable Model-Agnostic Explanations (LIME)

LIME is a post-hoc, model-agnostic explanation method that can reveal the influence of features on a single prediction by fitting a simple, interpretable linear model through local perturbations of the input data. LIME is a post-hoc, model-agnostic explanation technique that is able to disclose the feature contribution to an individual prediction by learning a simple, interpretable linear model via local perturbations of the input data. LIME is widely applied in text sentiment analysis, particularly for explaining deep models or hybrid methods, with the aim of making it clear for the user "which words/segments weighed most in the model's decision" in brief-text contexts like Twitter and product reviews (Lovera et al., 2021).

2. SHapley Additive exPlanations (SHAP)

SHAP, derived from Shapley values in game theory, calculates the marginal contribution of each input feature to the output result. Over the past several years, SHAP has been utilized to interpret the discrimination process of sophisticated models such as Transformer and BERT for sentiment analysis in particularly high-stakes fields such as finance and healthcare. It has been established that SHAP can be leveraged to generate "explainable lexicons" (e.g., XLex) automatically with great gain on interpretability and some degree of performance (Rizinski et al., 2024). SHAP is one of the most widely used interpretability techniques for text classification and sentiment analysis nowadays.

3. Attention Mechanism

Attention mechanism is not just one of the most significant technologies for enhancing model performance but also a natural approach for neural network interpretability. With the visualization of attention weights, it is possible to see "on which words/phrases the model paid attention" when discriminating sentiment, giving partial explainable hints for the model decision-making process. In recent years, a large body of research has been dedicated to improving the "fidelity" and "plausibility" of attention explanations, and has proposed enhancing the consistency between attention and human cognition through regularization (Tutek & Šnajder, 2022; Ghasemi & Momtazi, 2023).

4. Explainable Visualization

Regardless of whether it's LIME, SHAP, or the Attention mechanism, their results can ultimately be presented intuitively through visualization (e.g., heatmaps, keyword highlighting, feature weight maps), helping users understand the model's internal logic and increasing trust (Lovera et al., 2021).

2.5.3 Representative Applications of Explainable Methods in Sentiment Analysis

In practical sentiment analysis tasks, different explainable methods have been widely applied to various models and scenarios:

- Transformer and SHAP Integration: In financial sentiment analysis, researchers leveraged a Transformer model combined with SHAP to generate a domain-specific "explainable dictionary." This approach addressed the challenges of maintaining traditional manual dictionaries and improved both model performance and transparency (Rizinski et al., 2024).

- LIME for Explaining Deep Models: In Twitter sentiment analysis, LIME was employed as an interpretability tool for a hybrid model (knowledge graph + deep learning). It effectively

revealed the classifier's decision basis on individual samples, enhancing the model's traceability and user trust (Lovera et al., 2021).

- Exploring Attention Mechanism Interpretability: It has also been established that even if attention mechanisms can make decisions of neural nets understandable, their own interpretability must be enhanced through regularization and alignment according to human annotation so as not to be disrupted by "spurious attention distribution" (Tutek & Šnajder, 2022; Ghasemi & Momtazi, 2023).

- Explainable Multimodal Sentiment Analysis: Multimodal and cross-lingual sentiment analysis has also been enhanced by combining a number of techniques involving the integration of hybrid attention models and counterfactual explanations in an effort to enhance explainability in models. The model can be applied in a variety of low-resource languages and hard-to-reach scenarios (Ghasemi & Momtazi, 2023).

2.6 Misclassification Analysis and Evaluation of Multi-class Sentiment Modeling

2.6.1 Multi-class Misclassification Theory and Confusion Matrix

For multi-class sentiment analysis tasks, the confusion matrix represents a critical tool for observing and pinpointing misclassification effects. Confusion in multi-class classification is more complex compared to binary classification and happens in real application contexts in which intensity between opinions overlaps or boundaries between classes lose their distinctive quality. In a rating system where five stars will be given as a restaurant rating system, models confound rating scores between 4- and 5- or 3-stars. Confounding between adjacent classes can be visualized through the confusion matrix so as to pinpoint the vulnerabilities of the model precisely (Rizinski et al., 2024). The confusion matrix not only facilitates measurement of overall performance but also permits a nuanced analysis of each sentiment class identifiability and pitfalls common to each class.

Taking it a step further, the confusion matrix offers a stable basis for backtracking misclassifications and model refinement. For instance, examining regions of confusion between certain categories allows one to explore further the reasons for misclassification by taking into account sample text content and feature distribution. Some authors have suggested that the combination of confusion matrix with explainability analysis tools (such as LIME and SHAP) is able to better reconstruct the model's decision logic, offering a theoretical framework for the local and global optimization of sophisticated models (Lovera et al., 2021). In real business situations, the confusion matrix is also applied in conjunction with business objectives to identify 'critical is classifications" that most impact end decisions.

2.6.2 Types, Causes, and Evaluation Metrics of Misclassification

Types of misclassifications in multi-class sentiment classification consist primarily of nearby label misclassifications (e.g., 4-stars vs. 5-stars), extreme label confusions (e.g., 1-stars vs. 5-stars), and neutral vs. positive or negative polarity confusions. The causes of misclassifications tend to be based on the inherent subjectivity of data, ambiguous label meanings, highly imbalanced class distributions and complex contexts and words. Unless the model itself becomes sensitive enough toward boundary examples and minority events, it will also highly likely be misclassified as well. These issues notably manifest in real applications like fine-grained sentiment polarity classification and cross-platform multi-domain analysis (Ghatora et al., 2024).

For resolution of such misclassifications, the academic community uses a set of performance evaluation indices for the evaluation of a model's performance from a holistic viewpoint. In addition to extensively used Accuracy, Macro/Micro F1-score, Weighted F1, Recall, and Kappa coefficient, performance measurement in multiple aspects can also utilize ROC-AUC. In the specific situation of class imbalance and complex task intensity, concurrent examination of F1-score and confusion matrix can indicate clearly the sensitivity and error points of a model for each class and provide a scientific reference for subsequent adjustment of the model (Lovera et al., 2021; Han et al., 2020).

2.6.3 Typical Countermeasures and Frontier Methods

To reduce multi-class sentiment analysis misclassification rates, researchers have proposed various solutions both at data and model levels. At the data level, undersampling/oversampling, text augmentation, and pseudo-labeling are common practiced techniques to minimize classification bias due to class imbalance. At the model level, incorporating class-weighted loss functions, hierarchical classification model architectures, and model ensembles have significantly enhanced the ability of the model to distinguish minority classes and boundary samples. These methods have been widely shown to perform well on real-world multi-class tasks such as Yelp and Twitter (Lovera et al., 2021; Ghatora et al., 2024).

In the past few years, with the pace of explainable AI (XAI) technologies' development accelerating, researchers have begun applying explanation tools like LIME and SHAP to misclassification analysis and model diagnosis. By visualizing the decision-making of the model and emphasizing high-weight features, developers can identify semantic vulnerabilities in specific misclassified instances, and as a result, data annotation, feature engineering, and model architecture can be optimized in a targeted way (Ghasemi & Momtazi, 2023). Additionally, sophisticated techniques such as adversarial training and knowledge graph feature fusion are being used more and more to enhance the robustness and generalization capacity of multi-class models, opening the way for highly reliable applications of sentiment analysis systems in complex real-world environments (Tutek & Šnajder, 2022).

2.7 Research Gaps

Despite numerous advancements in the field of sentiment analysis, there remain several research gaps in multi-class sentiment modeling and explainability analysis based on large-scale user review data (e.g., Yelp):

First, the misclassification problem in multi-class fine-grained sentiment modeling remains unsystematically addressed. Existing models in multi-category tasks, such as 1–5 star ratings,

commonly exhibit phenomena like neighboring label confusion (e.g., 4-star vs. 5-star, 2-star vs. 3-star) and extreme misclassifications, which hinder practical business applications. Most mainstream research focuses on overall accuracy, with insufficient in-depth exploration of the causes of specific misclassification types, fine-grained visualization of confusion matrices, and the impact of typical confusions on subsequent decisions (Ghatora et al., 2024; Rizinski et al., 2024).

Second, the integration and application of misclassification explainability tools and methods in multi-class sentiment analysis are limited. Although explainability techniques like LIME, SHAP, and Attention have been used for single-sample or local explanations, there is currently a lack of mature engineering practices and case studies that combine these methods with multi-class misclassification analysis to form a systematic "misclassification attribution-optimization-visualization" process (Ghasemi & Momtazi, 2023; Lovera et al., 2021).

Third, there is insufficient research on explainability assisted by multi-source feature fusion and metadata. Current sentiment analysis models largely rely on primary textual features. Relevant research and empirical cases are relatively scarce concerning the synergistic explanation of structured metadata, such as user attributes, business categories, and temporal information, with text features. This makes it difficult for models to fully explain the causes of misclassifications and suggest optimization directions in real-world diverse scenarios (Rizinski et al., 2024).

Finally, for multi-class sentiment analysis on specific platforms (e.g., Yelp) facing real-world challenges such as label subjectivity and uneven data distribution, there is a lack of an explainability-driven misclassification analysis and system improvement framework. How to combine confusion matrices, local and global explanation methods, and diverse data features to achieve traceable and actionable model optimization remains a pressing research problem (Ghatora et al., 2024).

2.8 Summary

In summary, the field of sentiment analysis has made significant progress in areas such as dataset construction, algorithm optimization, explainability, and misclassification analysis. From the diversity of mainstream public datasets to the continuous evolution of traditional and deep learning methods, and the application and integration of explainability tools like LIME, SHAP, and Attention, existing research has laid a solid foundation for improving the performance and transparency of sentiment analysis models. However, for large-scale multi-class scenarios like Yelp, there are still prominent issues such as the complexity of misclassification types, the loose integration of explainability with misclassification analysis, and the insufficient synergistic explanation of multi-source features.

This chapter clarifies the main pain points and development bottlenecks in current sentiment analysis research through a literature review. It systematically sorts out typical methods and their advantages and disadvantages, and further summarizes the key directions to be broken through in the Research Gaps section. Subsequent chapters will focus on the research proposals and methodological innovations proposed to address the aforementioned gaps, striving to achieve high performance and strong explainability in multi-class sentiment modeling, and to provide strong support for practical business needs and academic development.

REFERENCES

Abo, M. E. M., Idris, N., Mahmud, R., Qazi, A., Hashem, I. A. T., Maitama, J. Z., ... & Yang, S. (2021). A multi-criteria approach for arabic dialect sentiment analysis for online reviews: Exploiting optimal machine learning algorithm selection. *Sustainability*, *13*(18), 10018.

Alamoudi, E. S., & Alghamdi, N. S. (2021). Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, *30*(2-3), 259-281.

Almansour, A., Alotaibi, R., & Alharbi, H. (2022). Text-rating review discrepancy (TRRD): an integrative review and implications for research. *Future Business Journal*, *8*(1), 3.

Arya, V., Mishra, A. K. M., & González Briones, A. (2022). Analysis of sentiments on the onset of COVID-19 using machine learning techniques.

Aslam, N., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model. *Ieee Access*, *10*, 39313-39324.

Benarafa, H., Benkhalifa, M., & Akhloufi, M. (2024). An Improved SVM Noise Tolerance for Implicit Aspect Identification in Sentiment Analysis. *Journal of Advances in Information Technology*, *15*(7).

Brandão, J. G., Junior, A. P. C., Pacheco, V. M. G., Rodrigues, C. G., Belo, O. M. O., Coimbra, A. P., & Calixto, W. P. (2025). Optimization of machine learning models for sentiment analysis in social media. *Information Sciences*, *694*, 121704.

Chatzimina, M. E., Papadaki, H. A., Pontikoglou, C., & Tsiknakis, M. (2024). A comparative sentiment analysis of Greek clinical conversations using BERT, RoBERTa, GPT-2, and XLNet. *Bioengineering*, *11*(6), 521.

Dandash, M., & Asadpour, M. (2025). Personality analysis for social media users using Arabic language and its effect on sentiment analysis. *Social Network Analysis and Mining*, *15*(1), 6.

Das, R. K., Islam, M., Hasan, M. M., Razia, S., Hassan, M., & Khushbu, S. A. (2023). Sentiment analysis in multilingual context: Comparative analysis of machine learning and hybrid deep learning models. *Heliyon*, *9*(9).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Fan, X., & Zhang, Z. (2024, July). A fine-grained sentiment analysis model based on multi-task learning. In *2024 4th International Symposium on Computer Technology and Information Science (ISCTIS)* (pp. 157-161). IEEE.

Fransisca, D., Sulistyowati, I., Budi, I., Santoso, A. B., & Putra, P. K. (2021, November). Sentiment Analysis of Office Automation Application in One of Indonesian Ministries. In *2021 5th International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 181-186). IEEE.

Ghasemi, R., & Momtazi, S. (2023). How a Deep Contextualized Representation and Attention Mechanism Justifies Explainable Cross-Lingual Sentiment Analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(11), 1-15.

Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., & Shaukat, N. (2024). Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM. *Big Data and Cognitive Computing*, *8*(12), 199.

Golubev, A. A., & Loukachevitch, N. V. (2021). Use of bert neural network models for sentiment analysis in russian. *Automatic Documentation and Mathematical Linguistics*, *55*, 17-25.

Han, K. X., Chien, W., Chiu, C. C., & Cheng, Y. T. (2020). Application of support vector machine (SVM) in the sentiment analysis of twitter dataset. *Applied Sciences*, *10*(3), 1125.

He, L. (2024). Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures. *Frontiers in Physics*, *12*, 1477714.

He, R., & McAuley, J. (2016, April). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web* (pp. 507-517).

Jain, R., Rai, R. S., Jain, S., Ahluwalia, R., & Gupta, J. (2023). Real time sentiment analysis of natural language using multimedia input. *Multimedia Tools and Applications*, *82*(26), 41021-41036.

Katić, T., & Milićević, N. (2018, September). Comparing sentiment analysis and document representation methods of amazon reviews. In *2018 IEEE 16th international symposium on intelligent systems and informatics (SISY)* (pp. 000283-000286). IEEE.

Kaur, P. (2022). Sentiment analysis using web scraping for live news data with machine learning algorithms. *Materials today: proceedings*, *65*, 3333-3341.

Khan, M. N., Khan, M. J., & Kashif, S. (2021). The Role of User Generated Content in Shaping a Business's Reputation on Social Media: Moderating role of trust propensity. *International Journal of Marketing, Communication and New Media*, *9*(16).

Lak, A. J., Boostani, R., Alenizi, F. A., Mohammed, A. S., & Fakhrahmad, S. M. (2024). RoBERTa, ResNeXt and BiLSTM with self-attention: The ultimate trio for customer sentiment analysis. *Applied Soft Computing*, *164*, 112018.

Lak, P., & Turetken, O. (2014, January). Star ratings versus sentiment analysis--a comparison of explicit and implicit measures of opinions. In *2014 47th Hawaii international conference on system sciences* (pp. 796-805). IEEE.

Lovera, F. A., Cardinale, Y. C., & Homsi, M. N. (2021). Sentiment analysis in Twitter based on knowledge graph and deep learning classification. *Electronics*, *10*(22), 2739.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NIPS), 30, 4765-4774.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142-150).

Nursal, A. T., Omar, M. F., Nawi, M. N. M., Khalid, M. S., Hanafi, M. H., & Deraman, R. (2025). Battle of Sentiment Lexicons: Wordnet, Sentiwordnet, Textblob and Vader in Web Forum Analysis. *Journal of Information Systems Engineering and Management*, *10*(2s).

Padhy, M., Modibbo, U. M., Rautray, R., Tripathy, S. S., & Bebortta, S. (2024). Application of Machine Learning Techniques to Classify Twitter Sentiments Using Vectorization Techniques. *Algorithms*, *17*(11), 486.

Qi, Y., & Shabrina, Z. (2023). Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. *Social network analysis and mining*, *13*(1), 31.

Rana, M. R. R., Nawaz, A., Ali, T., Alattas, A. S., & AbdElminaam, D. S. (2024). Sentiment Analysis of Product Reviews Using Transformer Enhanced 1D-CNN and BiLSTM. *Cybernetics and Information Technologies*, *24*(3), 112-131.

Rizinski, M., Peshov, H., Mishev, K., Jovanovik, M., & Trajanov, D. (2024). Sentiment analysis in finance: From transformers back to explainable lexicons (xlex). *IEEE Access*, *12*, 7170-7198.

Saudy, R. E., Alaa El Din, M., Nasr, E. S., & Gheith, M. H. (2022). A novel hybrid sentiment analysis classification approach for mobile applications Arabic slang reviews. *International Journal of Advanced Computer Science and Applications*, *13*(8).

Singh, A., Kalra, N., Singh, A., & Sharma, S. (2022, March). Sentiment analysis of Twitter data during Farmers' Protest in India through Machine Learning. In *2022 International Conference on Computer Science and Software Engineering (CSASE)* (pp. 121-126). IEEE.

Singh, N., & Jaiswal, U. C. (2023). Sentiment analysis using machine learning: A comparative study. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, *12*, e26785-e26785.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Tutek, M., & Šnajder, J. (2022). Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE access*, *10*, 47011-47030.

Wang, W. (2023). Sentiment Analysis: A Systematic Case Study with Yelp Scores. *Advances in Artificial Intelligence and Machine Learning. 2023; 3 (3): 74*. Machine Learning.

Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment analysis of Twitter data. *Applied Sciences*, *12*(22), 11775.

Wu, O., Yang, T., Li, M., & Li, M. (2020). Two-level LSTM for sentiment analysis with lexicon embedding and polar flipping. *IEEE Transactions on Cybernetics*, *52*(5), 3867-3879.