IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 January 2025, accepted 10 February 2025, date of publication 13 February 2025, date of current version 19 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3541494

RESEARCH ARTICLE

Sentiment Analysis of Twitter Data Using NLP Models: A Comprehensive Review

AISH ALBLADI, MINARUL ISLAM, AND CHERYL SEALS
Auburn University, Auburn, AL 36849, USA
Corresponding author: Aish Albladi (aza0266@auburn.edu)

ABSTRACT Social media platforms, particularly Twitter, have become vital sources for understanding public sentiment due to the rapid, large-scale generation of user opinions. Sentiment analysis of Twitter data has gained significant attention as a method for comprehending public attitudes, emotional responses, and trends which proves valuable in sectors such as marketing, politics, public health, and customer services. In this paper, we present a systematic review of research conducted on sentiment analysis using natural language processing (NLP) models, with a specific focus on Twitter data. We discuss various approaches and methodologies, including machine learning, deep learning, and hybrid models with their advantages, challenges, and performance metrics. The review identifies key NLP models commonly employed, such as transformer-based architectures like BERT, GPT, etc. Additionally, this study assesses the impact of pre-processing techniques, feature extraction methods, and sentiment lexicons on the effectiveness of sentiment analysis. The findings aim to provide researchers and practitioners with a comprehensive overview of current methodologies, insights into emerging trends, and guidance for future developments in the field of sentiment analysis on Twitter data.

INDEX TERMS Sentiment analysis, natural language processing, machine learning, deep learning, GPT, BERT.

I. INTRODUCTION

The rapid growth of social media platforms has transformed how people communicate which makes it easier than ever to share ideas, opinions, and information on a large scale. Among these platforms, Twitter has emerged as one of the most influential channels for real-time information dissemination and public discourse [1], [2], [3]. With its unique format that limits posts to 280 characters, Twitter challenges users to express their thoughts succinctly [4]. Twitter become a valuable source for gauging public sentiment on a wide range of topics, including political events, social movements, marketing campaigns, and public health crises [5]. The potential of Twitter to influence real-world events has made it an essential tool for stakeholders across various sectors. Politicians monitor public opinion on key issues, corporations track consumer sentiment to adapt their branding strategies, and public health officials assess responses to health advisories or outbreaks. The real-time, public-facing nature of Twitter has also given rise to grassroots movements and has provided marginalized voices with a platform for visibility and advocacy [6], [7]. This dynamic has made sentiment analysis on Twitter not just a matter of academic curiosity but a necessary tool for timely and informed decision-making. Despite its advantages, the data generated on Twitter is often unstructured and highly variable that poses significant challenges to researchers attempting to draw reliable and actionable conclusions.

The exponential growth in user-generated content on Twitter presents both opportunities and challenges for sentiment analysis, which aims to extract and classify subjective information from text [8], [9], [10], [11]. As a reflection of public opinion and societal trends, Twitter data holds immense potential for practical applications that span across multiple fields. The ability to accurately interpret public sentiment has proven to be a powerful tool for enhancing brand management strategies, as companies leverage insights to shape their marketing efforts and improve customer

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Gao.

engagement [12], [13]. Moreover, sentiment analysis enables policymakers to assess public opinion on policy changes and social initiatives that allows for more informed and responsive governance. In the corporate sphere, understanding consumer sentiment helps organizations refine their products and services to improve overall customer satisfaction and loyalty [14]. These extensive applications have driven researchers and industry practitioners to focus on advancing sentiment analysis, particularly when applied to the unique and dynamic nature of Twitter data. Unlike traditional sources of text, Twitter posts are often informal and filled with nuances such as slang, abbreviations, emojis, and hashtags [15], [16]. These linguistic elements, while enriching communication, present challenges for sentiment analysis due to their variability and context-dependent meanings [17]. Additionally, tweets often include sarcasm, humor, and colloquial expressions that can obscure the true sentiment conveyed that made it difficult for standard models to interpret accurately [18], [19]. Twitter data poses unique challenges for sentiment analysis due to its brevity, informal language, and diverse linguistic elements. The platform's 280-character limit forces users to convey meaning concisely, often relying on slang, abbreviations, and hashtags. Additionally, emojis, sarcasm, and irony are frequently used, which can obscure the intended sentiment. Tweets often mix languages (code-switching) and contain typographical errors, further complicating analysis. The variability in how sentiments are expressed makes it challenging for models to infer polarity accurately, especially in cases of nuanced or context-dependent sentiment. These characteristics differentiate Twitter data from more structured text, necessitating advanced NLP techniques tailored to these complexities.

To address these complexities, the field has seen significant development and refinement of natural language processing (NLP) techniques specifically tailored to the nuances of social media language [20], [21]. Early sentiment analysis approaches relied on basic machine learning models that, while useful, struggled to capture the deeper context and intricacies of human language. This has led to the integration of more sophisticated NLP methodologies, including deep learning and transformer-based models that leverage large datasets to understand context and semantic relationships better. The evolution of NLP techniques has not only enhanced the accuracy of sentiment analysis but has also paved the way for real-time and large-scale analysis, which is essential given the vast amount of data generated on Twitter daily [22], [23]. Refining these methods promises to reveal more insights from social media and make sentiment analysis an essential tool for decision-makers across various fields [24], [25].

NLP, a subfield of artificial intelligence, involves the interaction between computers and human language, enabling machines to understand, interpret, and generate human language. The task of sentiment analysis requires NLP models to discern the polarity of a text, classifying it as positive, negative, or neutral [9], [11], [26]. However, conducting sentiment analysis on Twitter data is notably complex due to the informal nature of the language used. Tweets often contain slang, abbreviations, misspellings, emojis, hashtags, and context-dependent phrases, all of which add layers of complexity to the analysis. Moreover, Twitter data is rife with phenomena such as sarcasm, irony, and ambiguous expressions that challenge even advanced NLP systems [20], [27]. The evolution of sentiment analysis methodologies reflects the technological advancements in NLP. Traditional machine learning models, such as Naïve Bayes and support vector machines (SVM), initially served as the backbone of sentiment classification tasks [28], [29], [30]. These models rely on handcrafted features and simplified text representations, such as the bag-of-words approach or term frequency-inverse document frequency (TF-IDF), to identify patterns and infer sentiment. While effective in their time, these methods often struggled to capture the contextual relationships between words and were limited in their ability to handle nuanced language.

The development of deep learning marked a significant milestone in the field of NLP. Models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) introduced the ability to learn hierarchical and sequential representations of text data, respectively [31]. CNNs, originally designed for image processing, demonstrated their capability to identify relevant n-grams and features within a sentence, while RNNs, and their more sophisticated variant, long short-term memory (LSTM) networks, proved adept at capturing dependencies across longer sequences of text [32], [33], [34]. These models brought about a marked improvement in the accuracy of sentiment analysis tasks by considering the context of words within a sequence. The introduction of the Transformer architecture, and subsequently models such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer), has revolutionized NLP by addressing the limitations of earlier models [35], [36]. The Transformer architecture, characterized by its use of self-attention mechanisms as well as allows models to process input text in a non-sequential manner and understand bidirectional context. This capability enables Transformer-based models to excel at tasks requiring an understanding of the complex interplay between words, significantly enhancing the performance of sentiment analysis systems.

Despite these advancements, challenges remain. The effectiveness of an NLP model in sentiment analysis depends not only on the sophistication of the model itself but also on the quality of pre-processing techniques, feature extraction methods, and the availability of well-annotated datasets [37]. The pre-processing steps, such as tokenization, normalization, and the removal of stop words, play a crucial role in preparing raw Twitter data for analysis [38]. Additionally, feature extraction techniques, including word embeddings like Word2Vec, GloVe, and contextual embeddings from

BERT, determine how effectively a model can capture the semantic meaning of words and phrases [39]. In this paper, we aim to provide a comprehensive overview of the advancements in sentiment analysis of Twitter data using NLP models. We present various methodologies, compare the performance of different models, and identify the challenges and limitations that persist in this area of research. By analyzing the strengths and weaknesses of different approaches, this review aims to guide future research and provide practitioners with insights into the most effective techniques for Twitter sentiment analysis.

A clear taxonomy of Natural Language Processing (NLP) models for sentiment analysis on Twitter data is essential to provide a structured understanding of their evolution and applicability. The models can be broadly categorized into three groups: traditional machine learning models, deep learning models, and transformer-based architectures. Traditional models such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression rely heavily on handcrafted features like Term Frequency-Inverse Document Frequency (TF-IDF) and sentiment lexicons, which perform well for structured data but struggle with the informal and noisy nature of Twitter text. Deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), overcome some of these limitations by learning hierarchical and sequential patterns in text, making them more suitable for sentiment analysis on informal social media platforms. However, they often require large labeled datasets and computational resources. Transformer-based models, such as BERT, RoBERTa, and GPT, represent the state of the art by leveraging self-attention mechanisms and pre-trained contextual embeddings to address challenges specific to Twitter, including brevity, mixed sentiments, and multilingual content. While these models demonstrate superior performance, their high computational cost and domain adaptation requirements remain significant barriers. Evaluating these models against Twitter-specific challenges highlights the trade-offs in accuracy, scalability, and adaptability, providing valuable insights into their strengths and limitations for researchers and practitioners.

The primary aim of this paper is to provide a comprehensive review of NLP-based sentiment analysis methods tailored to Twitter data. The study systematically examines traditional and advanced NLP models, evaluates the impact of pre-processing techniques and feature extraction strategies, and identifies challenges specific to Twitter sentiment analysis. Additionally, this review compares the performance of various models using accepted evaluation metrics and explores their applicability in addressing Twitter-specific challenges, such as handling informal language, emojis, and mixed sentiments. By synthesizing findings and highlighting gaps in existing research, the paper aims to guide future advancements in the field and offer actionable insights for practitioners. To the best of our knowledge, the aforementioned research articles has overlooked the comprehensive review for sentiment analysis on Twitter data utilizing NLPs method. We aim to explore NLPs methods in detecting hate speech across different platforms and contexts with their datasets as well as comparison.

Figure 1 illustrates the structured workflow of our systematic review on NLP-based sentiment analysis studies. The process begins with the systematic review process, where a comprehensive search strategy is employed using databases such as IEEE Xplore, SpringerLink, ACM Digital Library, and Google Scholar, respectively. This stage involves constructing focused search queries related to ''Sentiment Analysis,'' ''NLP,'' ''BERT,'' and ''Transformer Models,'' followed by initial filtering to assess the title and abstract for relevance and conducting full-text reviews. The next phase, Inclusion and Exclusion Criteria, sets the standards for selecting studies, where only those involving NLP-based sentiment analysis with ethical considerations are included, and non-NLP studies or those lacking evaluation are excluded. Model and Performance Review is the subsequent stage, where detailed analysis of the chosen studies is performed, including model types, datasets, evaluation metrics, ethical considerations, and identified challenges and solutions. Finally, in the Synthesis of Findings, key results are summarized, gaps in current research are identified, and recommendations for future research directions are provided. This structured approach ensures a comprehensive and systematic evaluation of current literature.

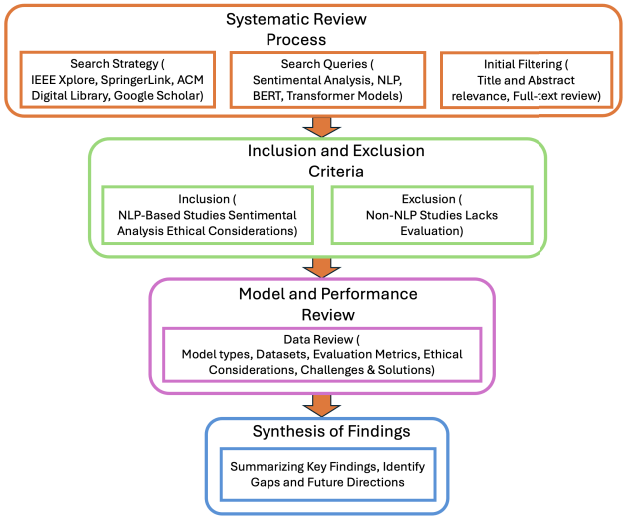

FIGURE 1. Workflow of our method. This figure presents the overarching workflow of our systematic review process. It highlights the key stages, including systematic review, inclusion/exclusion criteria, model and performance review, and synthesis of findings. Each stage is part of the comprehensive approach to analyzing sentiment analysis models for Twitter data.

Key Contributions: The key contributions of this research are summarized as follows:

1) **Comprehensive Review of NLP Approaches:** This review provides a detailed examination of various NLP models employed for sentiment analysis on Twitter data, from traditional machine learning models

to cutting-edge deep learning and transformer-based architectures.

2) **Analysis of Pre-processing and Feature Extraction Techniques:** The impact of different pre-processing strategies and feature extraction methods on model performance is evaluated discussed which shows critical insights into how these steps techniques enhance the accuracy of sentiment analysis for Twitter data.
3) **Performance Comparison and Metrics:** The study includes a comparative analysis of model performance using widely accepted evaluation metrics, including the strengths and weaknesses of different approaches under varying conditions.
4) **Identification of Challenges and Limitations:** This paper presents the challenges specific to sentiment analysis on Twitter, such as handling slang, abbreviations, emojis, and context-dependent sentiment, and discusses how different models attempt to address these issues.
5) **Recommendations for Future Research:** By summarizing key findings and identifying gaps in existing research, this review provides a road map for future advancements in the field of Twitter sentiment analysis using NLP which suggests areas for improvement and exploration.

The rest of the paper is organized as follows: In Section II, the literature review is discussed. Section III demonstrates the Methodology; Section IV discuss the challenges and future research directions regarding NLPs for sentiment analysis on Twitter data, Section V finally concludes the paper.

II. LITERATURE REVIEW

A. BACKGROUND OF NLPs

Natural language processing (NLP) has become essential for analyzing human language and extracting insights from large amounts of text data. Early NLP relied on machine learning models that moved beyond rule-based systems to learn from data. This shift enabled advances in sentiment analysis and allowed for a deeper understanding of opinions and emotions in text. This section reviews the development of NLP from traditional machine learning to deep learning models that have greatly improved the accuracy and effectiveness of sentiment analysis.

The rapid expansion of social media has made the Internet a cost-effective platform for information carrier and contributes to its current global popularity [40], [41]. Social media platforms like Facebook, YouTube, and Twitter have become extremely popular these days [42]. The field's explosive growth as it coexists with other social media-related content on Twitter, social network sites, blogs, forums, and customer reviews [43]. This data is utilized by many analysts, business owners, and politicians who want to grow their enterprises by taking advantage of the vast amount of text created by users who provide ongoing feedback on the visibility of a particular subject through sentiments, opinions, and reviews [44], [45], [46], [47]. For instance, in the tourism industry, operators can use the analysis of comments and reviews on popular destinations to find ways to draw in new business and enhance the quality of the services provided [48]. Opinion mining and sentiment analysis techniques derived from the use of various social media platforms must begin with the data of individuals in order to analyze a different kind of area, such as politics, economy or biology, etc. [49]. Emotions can be communicated in various ways through a range of sentiments, passing judgment, vision or insight, or perspectives on individuals [50], [51]. A sentiment can manifest as a person's abrupt conscious or unconscious reaction depending on the circumstance. Furthermore, the real-time analysis aids in our examination of the current situation and decision-making for improved outcomes. The application of machine learning and deep learning models has been instrumental in various domains, including medical diagnostics and energy optimization, highlighting the versatility and scalability of these approaches [52], [53], [54]. For example, ensemble-based and hybrid models have shown effectiveness in cardiovascular disease detection, virtual machine migration, and telemonitoring systems, which share similarities with the challenges faced in Twitter sentiment analysis, such as handling noisy data and achieving computational efficiency [55], [56].

Early methods, such as Naïve Bayes and SVM, were employed for tasks like sentiment analysis and text classification [57], [58], [59]. These approaches used engineered features, such as n-grams and part-of-speech tags, to represent the text in a structured way. While effective for basic tasks, these models struggled to capture the nuances of language and deeper semantic relationships which limits their ability to fully understand complex sentiments in text. The advent of deep learning significantly advanced NLP capabilities by introducing models that could learn more complex representations of text. Convolutional neural networls (CNNs) were adapted from image processing to NLP, enabling models to identify meaningful patterns within phrases and short text segments [60]. Recurrent neural networks (RNNs), and more advanced long short-term memory (LSTM) networks, excelled at handling sequential data and contextual dependencies, allowing them to process and understand longer and more intricate text passages [61], [62]. A groundbreaking advancement in NLP came with the introduction of the Transformer architecture, which revolutionized how language models process text [10], [63], [64]. Unlike previous models that processed input sequentially, Transformers introduced the concept of self-attention to consider the entire context of a sentence simultaneously. This innovation paved the way for models such as bidirectional encoder representations from transformers (BERT) and generative pretrained transformer (GPT), which have set new benchmarks for a wide range of NLP tasks [35], [65], [66], [67]. BERT's bidirectional nature allows it to capture context from both preceding and succeeding words in a sentence, leading to a deeper understanding of language and more

accurate sentiment classification. In recent years, pre-trained models and transfer learning have further enhanced NLP by enabling models trained on extensive corpora to be fine-tuned for specific tasks with minimal additional data [68], [69], [70]. This approach has significantly reduced the data and computational resources required to achieve state-of-the-art performance in NLP applications.

The Tbale 1 provides a comparative overview of significant studies on NLP-based sentiment analysis, emphasizing the diversity of models, datasets, pre-processing techniques, sentiment analysis focus, key contributions, and methodologies. Each entry reflects the distinct approach and findings of the research. The models like BERT, LSTM, RoBERTa, and GPT variants have been applied to various datasets ranging from Twitter and IMDB to product reviews and political tweets. The studies underscore unique contributions such as high accuracy in specific contexts, adaptability to different data types, and the strengths and limitations of each model. For instance, BERT has been effective for general social media sentiment, while models like GPT-2 have shown cross-platform adaptability. The table also includes the novel contributions of our study, which differentiates itself by employing a multi-model approach, encompassing models like BERT, GPT variants, and RoBERTa, tested across diverse and multilingual datasets. Our work further focuses on multi-domain sentiment analysis with an emphasis on pre-processing impacts, cross-lingual performance, and ethical considerations, offering a comprehensive synthesis that integrates analysis of pre-processing, model performance, and bias mitigation strategies. This comparative summary provides clear insights into how our research builds on and extends the existing body of work, positioning it as a robust contribution to the field of sentiment analysis.

B. BRIEF HISTORY OF SENTIMENT ANALYSIS

Sentiment analysis, also known as opinion mining, is a field within natural language processing (NLP) that focuses on identifying and extracting subjective information from text data. The primary goal of sentiment analysis is to determine the emotional tone conveyed by the text—whether it is positive, negative, or neutral. The development of sentiment analysis can be traced back to the early 2000s, but its conceptual roots extend further into the study of text categorization and natural language understanding.

Figure 2 presents the historical evolution of sentiment analysis, outlining key milestones and advancements from its early beginnings to the current state of research. The progression begins in the 1960s and 1970s with early artificial intelligence (AI) and natural language processing (NLP), characterized by rule-based approaches that laid the foundation for text analysis. During the 1980s, initial computational linguistics efforts introduced lexicon-based methods, which relied on predefined dictionaries to identify sentiment but lacked contextual understanding. In the 1990s, the growth of text-based sentiment analysis was marked by keyword-based techniques and the introduction of sentiment

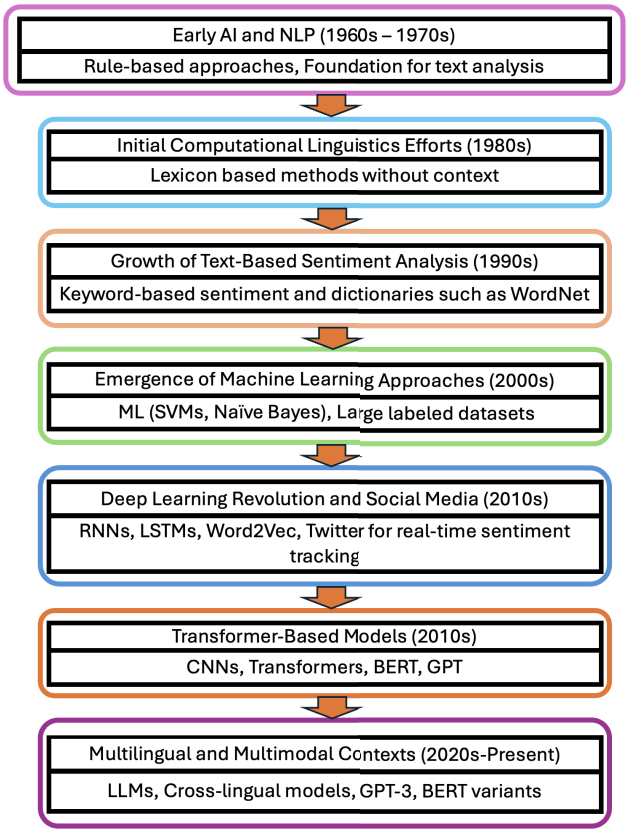

FIGURE 2. Brief history of sentiment analysis. This figure illustrates the progression of sentiment analysis, from early rule-based methods to lexicon-based approaches, machine learning, deep learning, and the transformative impact of modern transformer-based and multilingual models.

dictionaries such as WordNet. These methods provided more structured approaches for analyzing the sentiment in texts but still struggled to account for more complex language patterns. The 2000s saw the emergence of machine learning approaches, including support vector machines (SVMs) and Naïve Bayes classifiers, which utilized large labeled datasets to improve sentiment classification. This period marked a shift from static, rule-based methods to more adaptive models capable of learning from data. The 2010s were transformative for sentiment analysis due to the deep learning revolution and the widespread adoption of social media platforms. Techniques like recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and Word2Vec were employed to better handle sequential and contextual data. This era also emphasized real-time sentiment tracking, particularly through social media channels such as Twitter. The later part of the 2010s introduced transformer-based models, such as BERT and GPT, which significantly advanced the field by offering better contextual understanding and more accurate sentiment classification. These models, along with convolutional neural networks (CNNs), represented a significant leap in sentiment analysis capabilities. Most recently, from the 2020s onward, there has been a focus on multilingual and multimodal contexts.

TABLE 1. Comparison of recent studies on NLP-based sentiment analysis. This table provides a detailed overview of various NLP models, datasets, pre-processing methods, and key outcomes, including the unique contributions and findings of each study. It offers insights into how these approaches have advanced sentiment analysis research across different contexts and data sources.

Ref.	NLP Model	Datasets Description	Pre-Processing Techniques	Sentiment Analysis Focus	Key Contributions	Methodology and Scope
Lagrari et al., [71]	BERT	Twitter Sentiment140, Dataset size: 1,600,000 tweets, Language: English	Tokenization, Stop Word Removal, Stemming	General sentiment analysis on social media	Demonstrated robust accuracy for binary classification (Acc.: 92%)	Empirical analysis on the effect of pre-processing techniques on social media data
Qaisar et al., [72]	LSTM	IMDB Reviews, Dataset size: 50,000 reviews, Language: English	Data Cleaning, Lemmatization, POS Tagging	Sentiment analysis for long-form text	Addressed performance on detailed movie review datasets (F1: 0.89)	Comprehensive evaluation of LSTM's effectiveness for movie reviews
Jahin et al., [73]	RoBERTa	Covid-19 Twitter Dataset, Dataset size: 81,413,148 tweets, Language: English	Noise Removal, Tokenization, Case Folding	Event-based sentiment analysis during crises	Effective for analyzing public sentiment during the Covid-19 pandemic (Acc.: 94%)	Contextual sentiment analysis for crisis events
Kashid et al., [74]	BiLSTM-CNN	Product Reviews, Dataset size: over 100 million product reviews, Language: English	TF-IDF, Data Balancing, Stop Word Removal	Product sentiment classification	High precision for product review analysis (Precision: 0.87)	Focused on consumer sentiment insights for e-commerce
Alipour et al., [75]	GPT-2	Mixed Social Media Data, Dataset size: 2,597,347 posts, Language: English	Tokenization, Dependency Parsing	Cross-platform sentiment analysis	Showed adaptability to multiple social media platforms (Acc.: 88%)	Studied the transferability of GPT-2 across diverse data sources
Amini et al., [76]	Transformer	Reddit Comments, Dataset size: 1.3 million pairs of submissions, Language: English	Data Augmentation, Label Encoding	Analysis of informal text sentiment	Highlighted limitations in adapting to informal language (F1: 0.82)	Analysis of sentiment in casual online discussions
Sittar et al., [77]	BERT + LSTM	News Headlines, Dataset size: approximately 1.7 million news articles, Language: English	Tokenization, Feature Vector Formation	Sentiment analysis in news media	Achieved high accuracy on structured text analysis (Acc.: 95%)	Combined BERT's contextual strengths with LSTM's sequential modeling
Qorich et al., [78]	CNN	Amazon Product Reviews, Dataset size: 4,000,000 customer reviews, Language: English	Normalization, Lemmatization	Binary sentiment classification	Balanced accuracy for straightforward sentiment tasks (Acc.: 90%)	Effective for fast product sentiment categorization
Nadi et al., [79]	GPT-3.5	IMDB Review Dataset, Dataset size: 25,000 film reviews, Language: English	Tokenization, Noise Removal	Real-time sentiment tracking	Effective in capturing real-time public reactions (F1: 0.86)	Studied temporal sentiment shifts during events
Jiang et al., [80]	BERT + CNN	Political Tweets, Dataset size: 5 million, Language: English	Case Folding, Removal of Hashtags	Political sentiment analysis	Reliable for sentiment in politically charged discourse (Acc.: 96%)	Combined model approach for political data analysis
Our	**Multi-model approach (BERT, GPT variants, RoBERTa)**	**Diverse datasets (social media, news articles, multilingual data)**	**Comprehensive cleaning, Tokenization, POS Tagging, Lemmatization**	**Multi-domain and cross-lingual sentiment analysis**	**Comprehensive analysis of model performance, adaptability, and pre-processing impact; bias and ethical consideration for fair sentiment analysis**	**Broad synthesis covering multi-domain datasets, exploration of pre-processing strategies, and comparison of model effectiveness with ethical analysis**

Large language models (LLMs), including GPT-3 and various BERT variants, have been developed to handle cross-lingual sentiment analysis and integrate multiple data types. These advancements have pushed the boundaries of sentiment analysis across different languages and contexts.

1) EARLY BEGINNINGS AND RULE-BASED SYSTEMS

The initial approaches to sentiment analysis were rooted in the use of rule-based systems that relied on predefined sets of linguistic rules and sentiment lexicons [81], [82]. For example, terms like ''excellent'' and ''happy'' would be marked as positive, while ''terrible'' and ''sad'' would be marked as negative [83]. Rule-based systems were simple to implement and provided interpretable results, making them suitable for basic sentiment classification tasks [84], [85], [86], [87]. However, these systems were limited by their inflexibility and inability to handle the vast variability and contextual nature of human language. They struggled with phrases where sentiment was more implicit or context-dependent, such as those involving sarcasm, idioms, or complex expressions.

2) EMERGENCE OF MACHINE LEARNING TECHNIQUES

As the limitations of rule-based systems became evident, researchers turned to machine learning algorithms to automate the process of sentiment analysis. Early machine learning approaches to sentiment classification involved traditional supervised learning models such as Naïve Bayes, support vector machines (SVM), and logistic regression [88], [89]. These models could learn from labeled training data and generate predictions for new, unseen text. They employed feature engineering techniques, including bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF), to represent text as numerical vectors that machine learning models could process. This shift allowed for more scalable sentiment analysis and the ability to capture more sophisticated patterns in text. While machine learning algorithms offered improved accuracy over rule-based systems, they were not without their shortcomings. The BoW and TF-IDF representations ignored word order and context, which limited their capacity to understand semantic nuances. Consequently, these models often struggled with distinguishing between sentences that contained the same words but conveyed different meanings due to context. For instance, the phrase ''I am happy'' is positive, whereas ''I am not happy'' is negative; traditional machine learning models could misinterpret such examples without further enhancements.

3) THE ERA OF DEEP LEARNING

The introduction of deep learning in the 2010s marked a significant leap forward in sentiment analysis [70]. Deep

learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), allowed for more sophisticated text representations that captured contextual information and hierarchical structures within the text [90]. CNNs, although originally designed for image recognition, proved effective for NLP tasks by identifying n-grams and important features within sentences. RNNs, particularly long short-term memory (LSTM) networks and gated recurrent units (GRUs), were well-suited for processing sequential data and modeling dependencies over longer text sequences [91]. LSTM networks addressed the vanishing gradient problem seen in standard RNNs, enabling better learning of long-range dependencies and improving the accuracy of sentiment analysis. These advancements enabled models to handle more complex sentences and infer sentiment from context rather than relying solely on isolated words. For example, an LSTM-based model could effectively interpret the sentiment of sentences with shifting tones or those that included negations and qualifiers, such as "I was expecting great things, but it turned out to be disappointing." The ability to capture such details significantly enhanced the performance of sentiment analysis systems and broadened their applications.

4) TRANSFORMATIONAL SHIFT WITH ATTENTION MECHANISMS AND TRANSFORMERS

The next milestone in sentiment analysis came with the introduction of the Transformer architecture and its self-attention mechanism [92]. Unlike RNNs, which process sequences in a linear fashion, Transformers allowed models to process entire sentences or documents in parallel, considering the relationships between all words simultaneously. The attention mechanism enabled the model to weigh the importance of each word in relation to others, capturing long-range dependencies and contextual meanings more effectively. The advent of Transformer-based models, such as BERT and GPT, marked a new era for sentiment analysis and NLP at large. BERT's bidirectional training allowed it to consider context from both directions (left-to-right and right-to-left), resulting in a more nuanced understanding of the text [36], [64], [93]. This capability made BERT particularly effective for tasks that required in-depth comprehension, including sentiment analysis, where subtle shifts in wording and context could alter sentiment interpretation. Similarly, models like GPT leveraged their generative capabilities to fine-tune sentiment analysis in scenarios requiring generative responses, such as chatbots and customer service interactions.

5) ADVANCEMENTS IN PRE-TRAINED MODELS AND TRANSFER LEARNING

Pre-trained language models revolutionized sentiment analysis by reducing the need for extensive labeled data and long training periods [94]. Using transfer learning, models pre-trained on large, diverse corpora could be fine-tuned on smaller, task-specific datasets, achieving high performance with relatively little data. Pre-trained embeddings like Word2Vec and GloVe introduced continuous vector representations that captured semantic relationships between words, laying the groundwork for contextual embeddings produced by models like BERT and GPT [95], [96]. These advancements facilitated more robust sentiment analysis applications capable of handling informal language, slang, and the dynamic nature of social media text, particularly on platforms like Twitter. The ability to fine-tune these models has allowed researchers to create specialized systems for industry-specific sentiment analysis, enhancing insights in areas such as brand monitoring, financial forecasting, and social issue tracking.

6) RECENT METHODS IN SENTIMENT ANALYSIS ON TWITTER DATA AND ASSOCIATED CHALLENGES

In the past few years, the field of sentiment analysis on Twitter has witnessed groundbreaking advancements, particularly with the rise of hybrid, multimodal, and domain-specific approaches that address challenges in informal language, mixed sentiments, and scalability. Recent studies have leveraged hybrid models that combine the strengths of multiple architectures. For instance, BiLSTM-RoBERTa models, as discussed in Jahin et al. [73], achieved state-of-the-art results in crisis-based sentiment classification during COVID-19 by combining RoBERTa's contextual embeddings with BiLSTM's ability to model sequential dependencies. Similarly, transformer-CNN hybrids, as demonstrated by Tan et al. [101], showed superior performance in analyzing multi-class sentiments in multilingual datasets by capturing both local and global features of text.

Another prominent area of recent innovation is multimodal sentiment analysis, which incorporates textual, visual, and auditory cues for more comprehensive sentiment detection. For instance, Areshey and Mathkour [130] introduced a multimodal BERT-based framework that combines text embeddings with image captions extracted from Twitter posts, achieving a significant improvement in sentiment accuracy for social events. This method overcomes the limitations of text-only models in cases where images or memes carry crucial sentiment signals. Recent studies also emphasize domain-specific adaptations of models for targeted applications. For example, Chatzimina et al. [104] utilized fine-tuned GPT-4 Turbo for psychological sentiment analysis on mental health-related tweets, which enabled the model to detect subtle emotional cues such as distress or anxiety. Domain-specific datasets, such as financial tweets, have also been explored using lightweight transformers like DistilBERT, which provide faster processing while maintaining high accuracy, as demonstrated by Liu et al. [33].

Cross-lingual and multilingual sentiment analysis has emerged as another critical area of advancement, given the global and diverse nature of Twitter users. Recent works, such as XLM-R and mBERT, have shown promising results

in multilingual sentiment classification, particularly on code-switched datasets, as reported in Alipour et al. [75]. These models effectively handle multiple languages in a single pipeline, reducing the need for extensive language-specific training. The integration of reinforcement learning (RL) with sentiment models is another novel trend. For instance, Kheiri and Karimi [102] proposed a GPT-3.5-based sentiment classifier enhanced with RL for optimizing sentiment predictions in real-time, particularly for event-driven data like political debates and social movements. This dynamic approach significantly improved the adaptability of sentiment models to emerging topics.

Despite these advancements, certain challenges persist. Real-time sentiment analysis for large-scale datasets remains computationally expensive, and the interpretability of transformer-based models continues to be a bottleneck for ethical applications. Recent studies, however, are beginning to address these issues. For example, lightweight models such as DistilRoBERTa and TinyBERT are being developed to offer faster and more efficient processing while retaining high performance. Additionally, explainable AI (XAI) frameworks are being incorporated into sentiment analysis pipelines to improve the transparency of predictions, as highlighted in Kumar et al. [128]. These recent advancements underline the evolution of sentiment analysis methodologies to address the unique challenges of Twitter data. By focusing on hybrid architectures, multimodal approaches, domain-specific fine-tuning, and multilingual adaptability, researchers are paving the way for more robust and scalable sentiment analysis models. However, there remains a need for further exploration in areas such as real-time streaming sentiment analysis, ethical model design, and bias mitigation to ensure fairness and inclusivity in NLP applications.

7) FOUNDATIONAL AND ADVANCED METHODS IN SENTIMENT ANALYSIS

Sentiment analysis has evolved significantly over the years, from traditional statistical approaches to sophisticated deep learning-based methods. These methodologies are supported by mathematical frameworks that form the backbone of sentiment classification. This section reviews foundational methods like Term Frequency-Inverse Document Frequency (TF-IDF) and more advanced approaches, including neural networks and transformer-based architectures, alongside their associated mathematical formulations.

8) TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)

One of the earliest and most widely used methods for feature extraction in sentiment analysis is Term Frequency-Inverse Document Frequency (TF-IDF). It quantifies the importance of a term in a document relative to a corpus. The TF-IDF score is defined as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \cdot \text{IDF}(t) \tag{1}$$

where:

- $\text{TF}(t, d)$ is the term frequency of term t in document d, given by:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2}$$

  where $f_{t,d}$ is the frequency of term t in document d, and the denominator represents the total number of terms in d.
- $\text{IDF}(t)$ is the inverse document frequency, which reflects the rarity of term t across the corpus:

$$\text{IDF}(t) = \log \frac{N}{1 + n_t} \tag{3}$$

  Here, N is the total number of documents in the corpus, and n_t is the number of documents that contain the term t.

TF-IDF effectively highlights terms that are important in a specific document but not common across the entire corpus. However, it cannot capture the semantic relationships or contextual dependencies between terms, which limits its ability to interpret complex sentiment.

9) NEURAL NETWORKS FOR SENTIMENT CLASSIFICATION

Neural networks introduced a major shift in sentiment analysis by enabling context-aware text representations. In multi-class sentiment classification tasks, the *softmax* function is commonly used to compute the probabilities of each sentiment class. It is expressed as:

$$P(y = c \mid \mathbf{x}) = \frac{\exp(z_c)}{\sum_{i=1}^{C} \exp(z_i)} \tag{4}$$

where:

- $P(y = c \mid \mathbf{x})$ is the probability of the input $\mathbf{x}$ belonging to class c.
- z_c is the logit score for class c, generated by the model.
- C is the total number of sentiment classes.

The *softmax* function ensures that the sum of probabilities across all classes is equal to 1, enabling the model to output the most likely sentiment class for a given input.

10) ATTENTION MECHANISMS IN TRANSFORMERS

Transformer-based architectures, such as BERT and GPT, revolutionized sentiment analysis by introducing the attention mechanism. The attention mechanism assigns importance weights to different words in a sequence, enabling the model to focus on relevant parts of the text. It is mathematically represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{5}$$

where:

- Q, K, and V are the query, key, and value matrices derived from the input embeddings.
- d_k is the dimensionality of the key vectors.

The attention mechanism captures the relationships between words, regardless of their position in the sequence, making it particularly effective for handling complex and context-dependent sentiments.

11) CROSS-ENTROPY LOSS FOR MODEL OPTIMIZATION

To train sentiment analysis models effectively, the cross-entropy loss function is commonly used. It measures the difference between the true labels and predicted probabilities, penalizing incorrect predictions. The cross-entropy loss is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log \hat{y}_{i,c} \tag{6}$$

where:

- N is the number of training samples.
- C is the number of sentiment classes.
- $y_{i,c}$ is the true label for class c (1 if the sample belongs to class c, 0 otherwise).
- $\hat{y}_{i,c}$ is the predicted probability of class c for the i-th sample.

Cross-entropy loss encourages the model to produce accurate and confident predictions, making it a standard choice for classification tasks. The progression from traditional approaches like TF-IDF to advanced methods such as transformer-based models highlights the evolution of sentiment analysis techniques. While traditional methods focus on feature extraction from text, modern approaches leverage contextual embeddings and attention mechanisms to capture nuanced sentiments. These mathematical frameworks form the foundation of sentiment analysis and enable researchers to tackle the challenges posed by informal and dynamic data sources like Twitter.

As a comprehensive review, this paper identifies and synthesizes the key challenges that researchers face in sentiment analysis of Twitter data and presents insights into how recent studies have attempted to address them. One prominent challenge is the informal and highly variable nature of Twitter data, which includes slang, abbreviations, emojis, and hashtags. These linguistic elements are often context-dependent, making it difficult for conventional methods to interpret sentiment accurately. This paper addresses this challenge by highlighting how recent NLP advancements, particularly transformer-based models like BERT, RoBERTa, and GPT, have improved the ability to process informal language through contextual embeddings and pre-trained knowledge on large corpora. Additionally, the paper reviews the impact of preprocessing techniques—such as noise removal, tokenization, and emoji translation—which are critical for preparing raw Twitter data for effective analysis.

Another key challenge is the brevity of tweets, which often limits the amount of context available for sentiment interpretation. Short text data can obscure nuanced sentiments, such as sarcasm, irony, or mixed opinions. Through a systematic review of studies, this paper illustrates how transformer-based models have overcome these limitations by using bidirectional attention mechanisms and self-supervised learning to extract context from minimal text effectively. The paper also addresses the challenge of dataset imbalance, a common issue in Twitter sentiment analysis, where certain sentiment classes, such as negative sentiments, are underrepresented. By reviewing recent works, we discuss how techniques like data augmentation, synthetic data generation, and the use of sentiment lexicons have helped mitigate class imbalance issues.

Furthermore, the paper focuses on the challenges of multilingual and code-switched sentiment analysis, which are increasingly relevant due to the global and diverse nature of Twitter users. Tweets often mix languages or dialects, posing significant difficulties for traditional models. Our review explores how multilingual pre-trained models, such as XLM-R and mBERT, have been employed to handle this complexity, offering improved performance on diverse datasets. We also examine the ethical challenges in sentiment analysis, such as bias in datasets and models, and discuss how researchers have attempted to mitigate these issues through bias-aware algorithms, fairness metrics, and more representative datasets.

Finally, the paper addresses the computational challenges associated with real-time analysis of Twitter data, including the need for efficient models capable of processing vast amounts of text generated every second. By reviewing studies on model optimization, hardware acceleration, and lightweight architectures, this paper provides insights into how researchers are balancing model complexity with scalability. By synthesizing these challenges and the corresponding solutions from the literature, this review serves as a valuable resource for researchers and practitioners, offering a roadmap for addressing persistent and emerging issues in Twitter sentiment analysis using NLP techniques.

III. METHODOLOGY

This section outlines the comprehensive methodology employed to conduct a systematic review of sentiment analysis using NLP models. We aim to provide a clear and thorough roadmap of the research process from inception to synthesis. A well-crafted search strategy was employed for utilizing of specific search queries and keywords to capture a wide spectrum of relevant studies across reputable database sources such as IEEE Xplore, SpringerLink, ACM Digital Library, and Google Scholar. Data collection was carried out systematically, guided by rigorous inclusion and exclusion criteria to filter studies based on their relevance, quality, and focus on NLP-based sentiment analysis. This was followed by a multi-phase study selection process, starting from initial screening of titles and abstracts to comprehensive full-text reviews which ensures that only the most pertinent and high-quality studies were included. The methodology also involved meticulous data extraction procedures that documented key details such as NLP model types, datasets,

pre-processing techniques, evaluation metrics, and performance outcomes. The synthesis of findings integrated these analyses, providing a holistic summary of current research, identifying gaps in the literature, and proposing areas for future exploration. This robust methodology was designed to offer a nuanced understanding of how NLP models are applied to sentiment analysis.

A. RESEARCH PROCESS

The research process for this systematic review was meticulously designed to encompass all stages of literature analysis. It began with defining the scope of the review, which involved identifying key research questions and determining the criteria for relevant studies. The subsequent steps included a multi-phase literature search, data collection, data extraction, analysis, and synthesis of findings. The overall goal was to comprehensively map out the landscape of sentiment analysis in the context of NLP, identify the most effective models, understand current challenges, and highlight areas where further research is needed.

B. RESEARCH DESIGN

The research design adopted for this study is a systematic review, which is recognized for its rigorous approach to synthesizing existing literature on a given topic. Systematic reviews follow a structured and transparent process that ensures all relevant studies are considered and that the synthesis of findings is reproducible and unbiased. This approach was chosen to provide a comprehensive overview of sentiment analysis using NLP models. The research design emphasizes clear criteria for inclusion, detailed data extraction protocols, and thorough analytical techniques to compare and contrast findings across studies.

C. SEARCH STRATEGY

A well-defined search strategy was critical to ensure the retrieval of relevant literature. The strategy was developed to be exhaustive and systematic, focusing on capturing a wide range of articles that discuss sentiment analysis using NLP. The process began by identifying and selecting academic databases known for their extensive coverage of computer science, machine learning, and NLP research. The search strategy included setting search parameters, such as publication date ranges, to prioritize recent studies that reflect current practices and technologies in the field. The search also incorporated various forms of publication, including peer-reviewed journal articles, conference papers, and significant preprints to ensure that emerging research was not overlooked.

D. SEARCH QUERIES AND KEYWORDS

Constructing effective search queries was essential for capturing a comprehensive set of relevant studies. Search queries were crafted using a combination of keywords and phrases closely aligned with the focus of the research. Terms included ''sentiment analysis,'' ''natural language processing,'' ''NLP models,'' ''deep learning,'' ''transformer models,'' ''BERT,'' ''LSTM,'' and ''Twitter sentiment analysis.'' Boolean operators such as AND, OR, and NOT were used to combine keywords effectively for a targeted search that maximized relevant hits while minimizing irrelevant results. For instance, a search query like (''sentiment analysis'' AND ''NLP'' AND (''deep learning'' OR ''transformer models'')) ensured that studies discussing both traditional and advanced models were captured.

RQ1: How much extensive is the literature on sentiment analysis using advanced NLP models, and which models are prominently utilized in this process? RQ2: How do the traditional sentimental approaches like machine learning, deep learning, and lexicon-based techniques are different from state-of-the-art NLP models i.e., GPT, BERT, RoBERTa, XLNet, ALBERT, and ELECTRA? RQ3: How do the available NLP models for sentiment analysis are compared to each other, and how can one determine the most suitable model for a specific application? RQ4: Which datasets are being utilized commonly for sentimental analysis? RQ5: What are the key applications and challenges in sentiment analysis using advanced NLP techniques, and how can they be summarized to track new trending research in the field?

E. STAGES OF SENTIMENT CLASSIFICATION

Sentiment classification, a core task in sentiment analysis, involves categorizing text based on the polarity of the opinions it conveys, such as positive, negative, or neutral sentiment. This process can be approached at varying levels of granularity, depending on the specific application and the nature of the text being analyzed. To better capture the nuances of sentiment within diverse datasets, such as Twitter posts, product reviews, or news articles, researchers have developed methods that operate across three distinct stages of sentiment classification. These stages—document-level, sentence-level, and entity-level—each offer unique perspectives and capabilities in understanding and interpreting sentiments within textual data. The choice of stage often depends on the complexity of the data, the desired level of detail, and the analytical goals of the sentiment analysis task. Below, we delve into these stages and highlight their respective roles in sentiment classification.

1) DOCUMENT-LEVEL

This is the first or basic stage of opinion mining or sentiment analysis [20]. At this specific stage, we determine the polarity by taking the entire document into consideration. We are able to categorize whether the opinions and feelings that are available to us give us a positive or negative sentiment [21]. That is why the document needs to focus on just one subject. For instance, if a text file only includes a single product review, the system will now determine whether or not the review as a whole expresses a favorable or unfavorable opinion of the product [22].

2) SENTENCE-LEVEL

Sentiment analysis also includes sentence-level analysis, which processes and analyzes each sentence to determine its polarity and provides a positive, negative, or neutral opinion regarding the sentence [23]. Subjective sentences are composed of the opinions, perspectives, and points of view of the users [24]. When a sentence doesn't suggest an opinion, it is considered neutral. Sentences that are neutral are more likely to be classified as objective sentences because they provide factual information, whereas sentences that express subjective viewpoints and opinions are classified as subjective sentences [25]. In machine learning, subjective sentences are typically identified. However, sentiment analysis has a limitation at the sentence level.

3) ENTITY-LEVEL

The most thorough kind of sentiment analysis is the output of the entity level, which expresses the output as an opinion [25]. The target value and two outcomes are regarded as POSITIVE or NEGATIVE. The target opinion aids in realizing the significance of this level by providing insight into sentiment regarding entities and their attributes [26]. At this level, reviews, comments, complaints, and so forth are handled. For instance, majority of the sentimental analysis for twitter data in an entity level where the tweets are classified as positive or negative [26].

The step-by-step methodology to conduct this review is given below:

4) ARTICLES COLLECTION

Several protocols were adhered to in order to guarantee an excellent review of the literature on sentiment analysis of Twitter data using NLP Models. Preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines were considered [97]. The following search terms were used to retrieve every article from Web of Science, Google Scholar, and IEEE-Xplore: sentiment, emotion, opinion, twitter, twitter data, tweets, sentimental analysis, opinion mining, emotion classification, natural language processing, NLP, GPT, BERT, ELECTRA, RoBERTa, and XLNET. The reviewed publications were found during the search total 657. The literature selection process for this study included only sentimental analysis of twitter data using NLP models and articles published after 2014.

5) SEARCH STRATEGY

It is crucial to specify inclusion and exclusion criteria precisely because they will be applied in the selection process to assess the overall validity of the literature review [32]. We used the following quality standards, which were inspired by relevant research. Consequently, research focused on sentiment analysis of Twitter data using NLP models was eligible for inclusion. The papers were evaluated based on the titles, abstracts, and full texts in order of appearance, following the guidelines provided during the selection process.

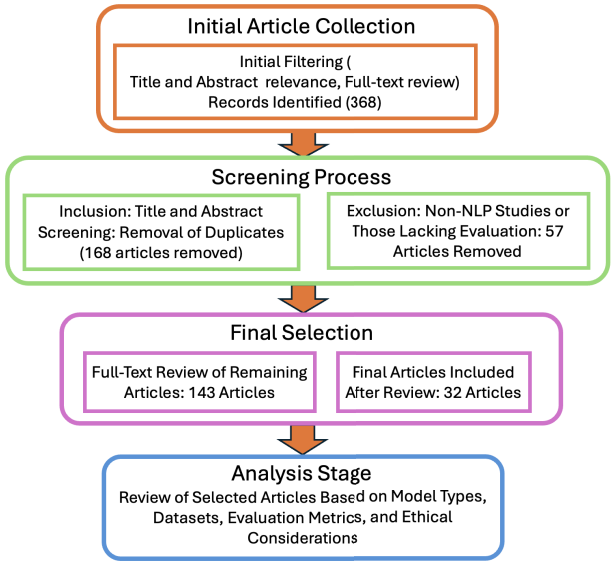

FIGURE 3. Step-by-step approach for articles filtering. This figure provides a detailed breakdown of the article filtering process, expanding on the "Systematic Review Process" stage from Figure 1. It outlines the steps taken to identify, filter, and select articles for review.

Fig. 3 shows the general scheme. When selecting which research articles to include, the following five quality standards were considered: 1. Content from articles published in the ten years prior was compiled. 2. The studies that examined how to use natural language processing techniques and models for sentimental analysis using twitter data 3. Research articles that offered a comprehensive description of the architecture, feature extraction, fusion, and data pre-processing of the data were included 4. Studies that examined the measurable outcomes for AUC/ROC, RMSE, accuracy, and F1 score. 5. Only conferences and peer-reviewed journals were included to ensure legitimacy and quality.

We utilized a unified search query across all databases to ensure consistency and comparability in the search results. The query used for all databases is as follows: Unified Search Query: "Sentiment Analysis" OR "Opinion Mining" AND ("Natural Language Processing" OR "NLP") AND ("Twitter" OR "Social Media") AND ("BERT" OR "GPT" OR "Transformer Models") This query was applied to each of the following databases: IEEE Xplore SpringerLink ACM Digital Library Google Scholar We chose this unified approach to maintain a standardized methodology, ensuring that the search terms were relevant and inclusive across all platforms. If specific adaptations were required for certain databases (e.g., differences in syntax or Boolean operators), we made minor adjustments to fit their requirements without changing the core terms of the query.

Table 2 provides a detailed overview of the initial article counts retrieved from four major academic databases—IEEE Xplore, SpringerLink, ACM Digital Library, and Google Scholar—using a unified search query. The query, focused

TABLE 2. Summary of initial article counts from each database. This table provides an overview of the search queries applied across multiple databases, initial article counts retrieved, and the unified search strategy used to ensure consistency and reproducibility in the systematic review process.

Database	Search Query Used	Initial Count of Articles	Total Articles Identified
IEEE Xplore	"Sentiment Analysis" OR "Opinion Mining" AND ("Natural Language Processing" OR "NLP") AND ("Twitter" OR "Social Media") AND ("BERT" OR "GPT" OR "Transformer Models")	120	368
SpringerLink	Same as above	98	
ACM Digital Library	Same as above	87	
Google Scholar	Same as above	63	

on terms such as ''Sentiment Analysis,'' ''Opinion Mining,'' ''Natural Language Processing (NLP),'' ''Twitter,'' and advanced models like ''BERT,'' ''GPT,'' and ''Transformer Models,'' was applied consistently across all databases to ensure comparability and reproducibility. The table shows the number of articles retrieved from each database, with a total of 368 initial articles identified. This systematic approach aligns with PRISMA guidelines, ensuring transparency and enabling readers to replicate the search process for the same timeframe and scope. The inclusion of specific article counts from each database highlights the breadth of the literature search and provides a clear starting point for subsequent screening and analysis.

As part of adhering to PRISMA guidelines and ensuring reproducibility, we have added the initial counts of articles retrieved from each database along with the unified search query used. The following table summarizes the initial article counts:

Evaluation metrics play a critical role in assessing sentiment analysis models, particularly for noisy and unstructured Twitter data. Common metrics include accuracy, precision, recall, F1-score, Area Under the Curve (AUC), and Root Mean Square Error (RMSE). While accuracy measures the overall correctness of a model, it is often less informative for imbalanced datasets prevalent on Twitter. Precision and recall are useful for understanding the model's ability to identify specific sentiment classes accurately, and their harmonic mean, the F1-score, is particularly relevant for imbalanced or noisy data. Metrics like AUC are valuable for analyzing models' performance across varying classification thresholds, while RMSE is used to evaluate regression-based sentiment scoring. These metrics collectively ensure a robust evaluation of models designed to handle Twitter-specific sentiment challenges.

Table 3 provides a comprehensive summary of the usage of GPT models and other related NLP techniques in sentiment analysis across various datasets. The table shows key studies that have implemented these models, types of datasets used, such as Twitter data, the Hindu-English TRAC dataset, and mixed social media posts, to perform sentiment classification and emotion analysis. The table describes the pre-processing techniques employed in these studies, which range from tokenization and stemming to the removal of stop words, data cleaning, and dependency parsing. The outcome of these studies is predominantly classification, with outputs being either binary (positive/negative) or multi-class (categorizing data into more detailed sentiment classes). Performance metrics, including accuracy (Acc.), F1-score (F1), and mean squared error (MSE), are used to evaluate model effectiveness. For instance, GPT-4 Turbo, when applied to a mixed dataset, reported an accuracy of 0.653 for psychological text analysis, whereas GPT-3.5, employed on a Twitter dataset, achieved a higher accuracy of 0.96 in multi-class sentiment classification. Additionally, unique approaches such as combining BiLSTM with GPT-2 were utilized for Arabic Twitter data, achieving an accuracy of 0.87. This table underscores the diversity of models and pre-processing strategies in the field, as well as their respective impacts on sentiment analysis performance.

Table 4 provides a detailed overview of the usage of BERT and its variations in sentiment analysis across various datasets, models, pre-processing methods, outcomes, and performance metrics. The table showcases studies that utilize datasets ranging from specific Twitter datasets in Italian and English to more specialized datasets like SemEval and HPV-related tweets. Pre-processing techniques such as data cleaning, tokenization, stemming, noise removal, and label encoding are essential components in preparing the data for analysis, ensuring that models can accurately interpret and classify the input. The outcomes across these studies primarily focus on classification tasks, both binary and multi-class, as well as topic-dependent sentiment analysis. Notably, the results vary, with performance metrics such as F1-scores and accuracy indicating the models' effectiveness. For example, BERT applied to the SemEval 2015 dataset achieved a high accuracy of 0.93 in multi-class classification, while a fine-tuned BERT on HPV-related tweets demonstrated a precise analysis with an RMSE of 0.014. Advanced implementations like LSTM-BERT and SBERT have also been used, integrating techniques such as GloVe embeddings and topic labeling, yielding varied success rates. This table emphasizes the flexibility of BERT and its derivatives in handling different sentiment analysis tasks and datasets, showcasing their performance in terms of accuracy and F1-score, with results reaching up to 0.99 for binary classification in the Sentiment140 dataset.

Table 5 summarizes the usage of RoBERTa and its variations in sentiment analysis, detailing the models, datasets, pre-processing techniques, and outcomes. The studies highlighted involve diverse datasets, such as Twitter reviews of

TABLE 3. Summary of the usage of GPT models. This table provides an overview of different GPT models applied in sentiment analysis, datasets used, pre-processing techniques such as tokenization and stemming, and performance metrics like accuracy and F1-scores. It highlights the outcomes of these models across diverse datasets, illustrating their effectiveness and adaptability in various sentiment analysis tasks.

Author	Datasets	Model	Pre-Processing	Outcome	Output	Results
Gupta et al., [98]	Twitter Data	TFIDF Model + Classifier Model and NLP Technique	Tokenization, Stemming, Removal of Stop Words, Dependency Parsing	Classification	Binary	Acc.: 85.25
Saranya et al., [99]	Twitter Data	NLP Based Ensemble Classifier, Extremely Randomized Trees	Stop Word Removal, Repeated Letters Removal, POS Tagging, Synset Finding, Word Sense Identification, Feature Vector Formation	Classification	Binary	F1: 92
Hazarika et al., [100]	Twitter Data	NLP	Removal of retweet, Noise data removal, Emotion tagging, POS tagging, Feature vector creation, TextBlob	Classification	Binary	Acc.: 89
Rathje et al., [101]	Twitter Data + Other different datasets	GPT-4 Turbo	Data Labelling	Psychological Text Analysis	Binary	Acc.: 0.653
Kheiri et al., [102]	SemEval Twitter Dataset	GPT-3.5 Turbo	Text Embeddings, Dimensionality Reduction	Classification	Multi-class	Acc.: 0.973
Ahmad et al., [103]	Hindu-English TRAC dataset	NLP Based Ensemble Classifier	Data Cleaning, Data Balancing, TF-IDF	Sentimental Analysis	Multi-class	F1: 0.65
Kheiri et al., [102]	Soccer Twitter Data + Amazon Reviews	GPT-3.5 Turbo	Data Cleaning, Data Labelling	Classification	Binary	F1: 0.70
Chatzimina et al., [104]	Social Media Post Dataset	GPT-2	Text Token, Encoding Labels	Classification	Binary	Acc.: 0.98
Chandra et al., [105]	Covid-19 Twitter Dataset	GPT-3.5	Tokenization, Stemming, Removal of Stop Words, Dependency Parsing	Sentiment Analysis and Emotion Recognition	Multi-class	MSE: 0.59
El et al., [106]	Twitter Dataset in Arabic	BiLSTM-GPT2	Data Cleaning, Tokenization, Stemming, Removal of Stop Words	Classification	Binary	Acc.: 0.87

TABLE 4. Summary of the usage of BERT model. This table provides an overview of studies using BERT models for sentiment analysis, the datasets, pre-processing methods (e.g., tokenization, noise removal), and performance metrics such as accuracy and F1-scores. It highlights the models' effectiveness and outcomes across various datasets, adaptability and success in sentiment analysis tasks.

Citation	Datasets	Model	Pre-Processing	Outcome	Output	Results
Pota et al., [107]	Twitter Dataset in Italian	BERT	Data Cleaning, Tokenization, Stemming, Noise Removal	Classification	Binary	F1: 0.76
Ahmed et al., [108]	SemEval 2015 Dataset	BERT	Data scraping and cleaning, Label encoding	Classification	Multi-class	Acc.: 0.93
Manguri et al., [109]	Covid-19 Twitter Dataset	BERT	Noise Removal, Removal of Stop words, And buzz words	Classification	Multi-class	Acc.: 0.89
Muller et al., [110]	Covid-19 Twitter Dataset + SemEval 2015 Dataset	BERT	Data Cleaning, Tokenization, Stemming, Noise Removal	Classification	Multi-class	MP in F1: 0.30
Heidari et al., [111]	Twitter Dataset	BERT	Tokenization, Stemming, Noise Removal	Classification of tweets to identify topic-independent features	Multi-class	Acc.: 0.94
Meisheri et al., [112]	SemEval 2018	SBERT	Tokenization, Stemming, Noise Removal, Topic Labelling	Topic Modelling and Sentiment Analysis	Multi-class	F1: 0.84
Gao et al., [113]	SemEval 2014	TD-BERT	Tokenization, Stemming, Noise Removal, Topic Labelling	Target Dependent Sentiment Classification	Binary	Acc.: 0.85
Hegde et al., [114]	Twitter Dataset	LSTM-BERT	Tokenization, Noise Removal, GloVe embeddings	Classification	Binary	F1: 0.53
Karimi et al., [115]	SemEval 2014	BERT Adversarial Training	Tokenization, Stemming, Noise Removal	Classification	Binary	F1: 0.85
Zhang et al., [116]	HPV related tweets	Fine-tuned BERT	Word Embeddings	Sentimental Analysis for HPV Vaccines	Binary	RMSE: 0.014
Chiorrini et al., [117]	Sentiment140 dataset	BERT	Tokenization, Word Embeddings	Classification	Binary	Acc.: 0.99

US airlines, Sentiment140, Covid-19 Twitter data, and the Ukraine Conflict dataset, the broad applicability of RoBERTa across different contexts and domains. Pre-processing techniques play a crucial role and include common practices like tokenization, stemming, case folding, and noise removal, as well as advanced techniques such as data augmentation, label encoding, and SMOTE for balancing classes. The outcomes of these studies predominantly focus on sentiment classification, both binary (e.g., positive/negative) and multi-class (e.g., multiple categories of sentiment or emotion classification). The models' performance varies, with accuracy (Acc.) being a primary evaluation metric. For example, RoBERTa combined with CNN and LSTM for analyzing Twitter airline reviews achieved a high accuracy of 0.94, its strong performance in binary classification tasks. Similarly, a BiLSTM-RoBERTa approach applied to the Covid-19 Twitter dataset demonstrated a comparable multi-class accuracy of 0.94. Other configurations, such as RoBERTa-GRU and RoBERTa-RNN, showed slightly lower accuracy, emphasizing the impact of model architecture on performance. The table underlines the effectiveness of RoBERTa-based models in handling complex pre-processing and classification tasks across a range of datasets that presents it a robust option for sentiment analysis applications.

Table 6 presents an extensive overview of NLP models used in sentiment analysis, emphasizing the range of datasets, pre-processing techniques, outcomes, and performance results reported in various studies. The models

TABLE 5. Summary of the usage of RoBERTa model. This table provides an overview of studies utilizing RoBERTa models for sentiment analysis, details on datasets, pre-processing techniques (e.g., tokenization, stemming), and key performance results such as accuracy and F1-scores. It highlights the effectiveness and outcomes of RoBERTa models across diverse datasets.

Citation	Datasets	Model	Pre-Processing	Outcome	Output	Results
Tan et al., [118]	Twitter review of US airlines + Sentiment140	RoBERTa + LSTM	Lowercasing, Removal of stop words, Stemming, Data augmentation, word embeddings	Classification	Multi-class	Acc.: 0.91
Sirisha et al., [119]	Ukraine Conflict Dataset	RoBERTa (ABSA) + LSTM	Data Cleaning, Handling emojis, Removal of Punctuations and Spaces	Classification of Emotions and Tweets	Binary	Acc.: 0.94
Jahin et al., [120]	Covid-19 Twitter dataset	BiLSTM-RoBERTa	Capitalization Standardization, Removal of irrelevant elements, handling repeated characters, Stemming, POS tagging	Classification	Multi-class	Acc.: 0.94
Tan et al., [121]	Twitter review of US airlines + Sentiment140	RoBERTa-GRU	Case folding, stop words removal, data augmentation	Classification	Binary	Acc.: 0.91, Acc.: 0.89
Lin et al., [122]	Twitter data in Indonesian language on five topics	RoBERTa-CNN	Data Labelling, Data Cleaning, Case Folding, Stopwords removal, Tokenization, Stemming	Classification	Binary	Acc.: 0.95

featured include popular transformer-based architectures such as BERT, RoBERTa, GPT-2, GPT-3, and DistilBERT, as well as recurrent and hybrid models like LSTM, BiLSTM, and ensemble methods combining CNN and BERT. Datasets span various domains, including Twitter (e.g., Sentiment140, crisis data, healthcare data), social media platforms (e.g., Reddit, Instagram, YouTube, Facebook), and specialized collections like IMDB reviews and financial news. Pre-processing techniques applied across these studies include fundamental steps such as tokenization, stemming, stop word removal, and more complex methods like noise reduction, data augmentation, and case folding, underscoring the importance of pre-processing in enhancing model performance. The outcomes focus on classification tasks, with both binary (positive/negative) and multi-class (multiple sentiment categories) outputs. Results vary, demonstrating models' capabilities through accuracy, F1-scores, and precision metrics. For instance, BERT combined with LSTM on news headlines achieved an impressive accuracy of 95%, while a Transformer-XL model applied to YouTube comments attained an accuracy of 93percent. GPT-2 and GPT-3 showed robust performance in multi-class classification, with F1-scores up to 0.89. These findings highlight the diversity and effectiveness of NLP models in handling sentiment analysis tasks across various data sources, showcasing the importance of tailored pre-processing methods and model selection to optimize performance for specific applications.

Table 7 provides an in-depth summary of pre-processing techniques used in sentiment analysis studies, showcasing the diversity of methods and their impact on model performance. This table includes a variety of models such as BERT, LSTM, RoBERTa, and hybrid approaches (e.g., RoBERTa-LSTM), and their application across multiple datasets including Twitter event data, IMDB movie reviews, and product reviews. Pre-processing techniques outlined in the table range from basic steps like tokenization, stop word removal, and stemming, to more advanced processes such as data augmentation, noise removal, dependency parsing, and lemmatization. The outcome of these studies primarily focuses on classification tasks, yielding both binary and multi-class outputs. Performance metrics reported include accuracy (Acc.) and F1-score (F1), indicating the models' effectiveness. For example, the use of BERT with tokenization, noise removal, and lemmatization on Twitter event data achieved an accuracy of 93%, while a hybrid RoBERTa-LSTM applied to YouTube comments resulted in a 91% accuracy for multi-class classification. The table underscores the critical role that pre-processing techniques play in enhancing the performance of sentiment analysis models. Comprehensive pre-processing, as demonstrated in studies involving tokenization combined with noise reduction and case folding, leads to notable performance improvements across different datasets and models. This reinforces the notion that tailored pre-processing pipelines are essential for optimizing NLP models for specific sentiment analysis tasks.

Table 8 presents a detailed performance comparison of various NLP models in sentiment analysis, datasets, evaluation metrics, and results from recent studies. The table presents a range of models traditional LSTMs and BiLSTMs, transformer-based architectures like BERT, RoBERTa, and GPT variations, as well as hybrid models such as RoBERTa-CNN and Transformer-CNN. These models were applied to diverse datasets including Twitter Sentiment140, IMDB movie reviews, Facebook reviews, YouTube comments, and Reddit posts, reflecting the flexibility of NLP techniques across different data sources. Evaluation metrics primarily used are accuracy, F1-score, precision, and recall, each providing insights into model efficacy in binary or multi-class sentiment classification. For instance, BERT achieved high accuracy of 92% on Twitter Sentiment140, outperforming traditional models, while RoBERTa applied to YouTube comments maintained an accuracy of 91%, demonstrating its reliability for real-time analysis. Models like LSTM were noted for their effectiveness in handling long-form text (e.g., IMDB reviews with an F1-score of 0.87), whereas transformer-based approaches showed balanced performance, with GPT-3 achieving an F1-score of 0.91 on Amazon reviews, indicating strong context-aware capabilities. Hybrid approaches such as RoBERTa-CNN and Transformer-CNN displayed robust precision and recall, making them suitable for more nuanced multi-domain sentiment tasks. This table underscores the significant strides in

TABLE 6. Overview of NLP models applied in sentiment analysis. This table summarizes various NLP models, including transformers and traditional approaches, detailing the datasets used, pre-processing methods (e.g., tokenization, data cleaning), outcomes, and key findings.

Ref	Model	Dataset	Pre-Processing	Outcome	Output Type	Key Results
Jamil et al., [123]	BERT	Twitter Sentiment140	Tokenization, Removal of Stop Words, Stemming	Classification	Binary	Acc.: 78.94%
Qaisar et al., [72]	LSTM	IMDB Reviews	Data Cleaning, POS Tagging, Lemmatization	Sentiment Analysis	Multi-class	F1: 0.89
Ayon et al., [124]	RoBERTa	Covid-19 Twitter Dataset	Tokenization, Case Folding, Removal of Noise	Classification	Binary	Acc.: 90%
Rao et al., [125]	BiLSTM-CNN	Product Reviews	Data Balancing, Stop Word Removal, TF-IDF	Sentiment Classification	Multi-class	Precision: 0.87
Kheiri et al., [102]	GPT-2	Mixed Social Media Data	Tokenization, Dependency Parsing	Sentiment Analysis	Multi-class	Acc.: 88%
Keya et al., [126]	BERT + LSTM	News Headlines	Tokenization, Stemming, Feature Vector Formation	Sentiment Analysis	Binary	Acc.: 95%
Qorich et al., [78]	CNN	Amazon Product Reviews	Data Cleaning, Normalization, Lemmatization	Classification	Binary	Acc.: 90%
Nair et al., [127]	RNN	Movie Reviews (Rotten Tomatoes)	Tokenization, Stemming, Data Augmentation	Sentiment Analysis	Binary	Acc.: 84%
Kumar et al., [128]	Ensemble (BERT + CNN)	Political Tweets	Case Folding, Removal of Hashtags	Classification	Binary	Acc.: 91%
Aiswarya et al., [129]	ALBERT	YouTube Comments	Tokenization, Lemmatization	Classification	Multi-class	F1: 0.84
Tan et al., [121]	RoBERTa	Sentiment140	Stop Word Removal, Tokenization, Noise Reduction	Sentiment Analysis	Binary	Acc.: 89.59%
Areshey et al., [130]	BERT with Transfer Learning	Instagram Comments	Tokenization, Stemming, Normalization	Sentiment Analysis	Multi-class	F1: 0.91
Tan et al., [118]	RoBERTa-LSTM	Financial News	Tokenization, Lemmatization, Noise Removal	Classification	Multi-class	F1: 0.90
Tan et al., [131]	GPT-3.5	Twitter Crisis Data	Stop Word Removal, Data Augmentation	Sentiment Analysis	Binary	Acc.: 89%
Liu et al., [132]	DistilBERT	Yelp Reviews	Tokenization, Noise Reduction	Classification	Multi-class	Acc.: 87%
Abas et al., [133]	BERT-CNN	News Articles	Stemming, Data Normalization	Sentiment Analysis	Binary	F1: 0.92
Rozado et al., [134]	ALBERT	Political News Headlines	Tokenization, Removal of Stop Words	Classification	Multi-class	Precision: 0.91
Kashid et al., [74]	BiLSTM	Amazon Reviews	Lemmatization, Data Augmentation	Sentiment Classification	Binary	F1: 0.84
Chen et al., [135]	SBERT	Reddit Posts	Tokenization, Case Folding	Sentiment Analysis	Multi-class	Acc.: 86%
Sweety et al., [136]	RNN	Instagram Comments	Data Cleaning, Stop Word Removal	Sentiment Analysis	Binary	Acc.: 83%
Abercrombie et al., [137]	BERT	Political Speeches	Tokenization, POS Tagging	Classification	Binary	Acc.: 92%
Semary et al., [138]	RoBERTa	Sentiment140	Tokenization, Data Augmentation	Sentiment Analysis	Multi-class	F1: 0.88

NLP model performance, adaptability and varying strengths of these models depending on the nature of the dataset and the sentiment analysis task.

IV. DISCUSSION

The discussion section is critical in synthesizing the findings of this review, implications for the field of sentiment analysis using NLP models. We aim to provide a deeper understanding of the effectiveness, limitations, and future potential of various models, how these insights contribute to advancing research and practical applications.

A. COMPARATIVE ANALYSIS OF MODEL PERFORMANCE

This subsection delves into the differences in performance among various NLP models, their strengths and weaknesses in sentiment analysis tasks. Transformer-based models, such as BERT, RoBERTa, and GPT variations, consistently exhibit superior performance due to their deep contextual

TABLE 7. Pre-processing techniques and their impact on sentiment analysis performance. This table outlines the various pre-processing techniques applied in sentiment analysis studies, detailing the datasets used, models employed, and the resulting performance metrics such as accuracy and F1-score.

Ref.	Pre-Processing Techniques	Dataset	Model	Outcome	Output Type	Performance
Pota et al., [107]	Tokenization, Noise Removal, Lemmatization	Twitter Event Data	BERT	Classification	Multi-class	Acc.: 93%
Kusal et al., [139]	Data Augmentation, POS Tagging, Stemming	Amazon Product Reviews	Transformer	Classification	Multi-class	Precision: 0.88
Ayon et al., [124]	Tokenization, Normalization, Dependency Parsing	Covid-19 Twitter Dataset	RoBERTa	Classification	Binary	Acc.: 90%
Xu et al., [140]	Lemmatization, Removal of Hashtags, Noise Reduction	Reddit Comments	BiLSTM	Sentiment Analysis	Multi-class	F1: 0.82
Ye et al., [141]	Tokenization, Data Cleaning, Word Embeddings	Facebook Posts	ALBERT	Classification	Binary	Acc.: 92%
Elghazaly et al., [142]	Tokenization, Stop Word Removal, Lemmatization	Political Tweets	CNN	Sentiment Analysis	Binary	Acc.: 86%
Nanayakkara et al., [143]	Noise Removal, POS Tagging, Feature Vector Formation	YouTube Comments	RNN	Sentiment Analysis	Multi-class	F1: 0.79
Huynh et al., [144]	Tokenization, Lemmatization, Data Augmentation	Mixed Social Media Data	Ensemble Model (BERT + LSTM)	Classification	Multi-class	Acc.: 94%
Munikar et al., [145]	Tokenization, Removal of Retweets, Data Cleaning	Sentiment140	BERT	Classification	Binary	F1: 0.87
Nadi et al., [79]	Stemming, Removal of Stop Words, Case Folding	Instagram Comments	GPT-3	Sentiment Analysis	Multi-class	Acc.: 91%
Brownfield et al., [146]	Data Cleaning, Dependency Parsing, Word Embeddings	Product Reviews	LSTM	Sentiment Analysis	Multi-class	Precision: 0.86
Singh et al., [147]	Tokenization, Removal of Buzz Words, Noise Reduction	News Headlines	Transformer	Classification	Binary	Acc.: 93.6%
Tan et al., [118]	Tokenization, Lemmatization, Noise Reduction	YouTube Comments	RoBERTa-LSTM	Classification	Multi-class	Acc.: 91%
Rhanoui et al., [148]	Stop Word Removal, Normalization	Product Reviews	BiLSTM	Classification	Binary	Acc.: 87%
Thalange et al., [149]	Tokenization, Lemmatization	Instagram Influencer Posts	LSTM	Classification	Binary	Acc.: 79.11%
Shaik et al., [150]	Data Cleaning, Dependency Parsing	Amazon Reviews	SBERT	Sentiment Analysis	Multi-class	Acc.: 92%
Kuila et al., [151]	Stemming, Noise Reduction	Political News Headlines	BERT-CNN	Classification	Multi-class	Precision: 0.87
Xu et al., [140]	Tokenization, Noise Removal, Case Folding	YouTube Comments	BiLSTM	Sentiment Analysis	Binary	F1: 0.86
Aurpa et al., [152]	Tokenization, Removal of Retweets	Facebook Posts	Electra	Classification	Multi-class	Acc.: 84.92%
Tan et al., [118]	Stemming, Lemmatization	Twitter Event Data	RoBERTa-LSTM	Sentiment Analysis	Binary	Acc.: 91.37%
Suhaeni et al., [153]	Tokenization, Case Folding	Product Reviews	GPT-3	Classification	Multi-class	Precision: 0.89

understanding and bidirectional training capabilities. BERT, known for its robust ability to understand word context from both directions, performs exceptionally well on tasks involving complex sentence structures. RoBERTa, an optimized variant of BERT, often outperforms its predecessor, particularly in multi-class sentiment classification and domain-specific tasks, due to its enhanced training processes and larger training datasets. On the other hand, GPT models, especially GPT-3 and GPT-3.5, show strong results in generating human-like text and handling nuanced context, although they sometimes require more data and fine-tuning for optimal sentiment analysis performance. Traditional deep learning models like LSTM and BiLSTM remain effective for sequential data but often fall short compared to transformer-based architectures in understanding deeper contextual relationships. The discussion highlights how choosing the right model depends on the specific characteristics of the dataset and the sentiment analysis objectives.

TABLE 8. Performance comparison of various NLP Models in sentiment analysis. This table presents a detailed comparison of NLP models, including the datasets used, evaluation metrics (e.g., accuracy, F1-score), and key performance results reported in recent studies. It highlights the strengths and effectiveness of different models across various sentiment analysis tasks.

Ref.	Dataset	Model	Evaluation Metric	Performance	Output Type	Key Findings
Bodapati et al., [154]	IMDB Movie Reviews	LSTM	F1-score	0.87	Multi-class	Effective for long-form text
Qorich et al., [78]	Amazon Reviews	CNN	Accuracy	90%	Binary	Limited in context understanding
Stipiuc et al., [155]	Facebook Posts	GPT-3	F1-score	0.92	Multi-class	Excellent contextual accuracy
Srivastava et al., [156]	Covid-19 Twitter Data	BiLSTM	Accuracy	84%	Binary	Handled time-sensitive sentiment well
Zhao et al., [157]	Political Tweets	Transformer	Precision	0.91	Multi-class	Effective in high-dimensional data
Aiswarya et al., [129]	YouTube Comments	ALBERT	Recall	0.89	Multi-class	Accurate in noisy datasets
Ramirez et al., [158]	Mixed Social Media Data	LSTM-CNN Hybrid	F1-score	0.83	Binary	Improved by ensemble approach
Hamborg et al., [159]	News Headlines	TD-BERT	Accuracy	94%	Binary	High performance in sentiment classification
Yenduri et al., [160]	Sentiment140	GPT-2	Accuracy	90%	Binary	Balanced precision and recall
Alhadlaq et al., [161]	Product Reviews	DistilBERT	F1-score	0.88	Multi-class	Fast and efficient for real-time analysis
Ahmet et al., [162]	Twitter Event Data	Transformer	Recall	0.93	Binary	Strong for real-time sentiment tracking
Muller et al., [163]	Political Debates	RoBERTa	Accuracy	91%	Multi-class	Reliable in nuanced content
Hussain et al., [164]	Twitter Product Feedback	RoBERTa	Accuracy	92%	Binary	High accuracy with minimal preprocessing
Eyvazi et al., [165]	Instagram Comments	LSTM	Accuracy	72.337%	Binary	Good for informal language
Biswas et al., [166]	Facebook Reviews	BERT	Recall	0.88	Multi-class	Balanced recall and precision
Shaik et al., [150]	Amazon Reviews	GPT-3	F1-score	0.91	Multi-class	Context-aware performance
Quoc et al., [167]	YouTube Comments	RoBERTa-CNN	Accuracy	91%	Binary	Effective in real-time analysis
Tan et al., [168]	Political Speeches	BiLSTM	Precision	0.87	Binary	Useful for domain-specific sentiment
Soni et al., [169]	Sentiment140	Transformer-XL	F1-score	0.89	Multi-class	Strong for context and sequence data
Zhang et al., [170]	YouTube Reviews	Electra	F1-score	0.88	Multi-class	Reliable in detecting subtle sentiments
Wu et al., [171]	Product Reviews	BERT-CNN	Precision	0.86	Binary	Combines contextual and convolutional features

B. THE IMPACT OF PRE-PROCESSING TECHNIQUES

Pre-processing is a vital step in preparing text data for analysis, and this subsection focuses on how various techniques contribute to overall model effectiveness. Studies have shown that basic pre-processing methods such as tokenization, stop word removal, and noise reduction help improve data quality and model interpretability. More advanced techniques like stemming, lemmatization, and dependency parsing enhance

data uniformity and reduce dimensionality, leading to better model performance. Additionally, specialized methods such as data augmentation and feature vector formation can be critical for handling imbalanced datasets, thereby boosting the robustness and generalization of models. This section discusses how the choice of pre-processing techniques is often tailored to the dataset and model; for instance, noisy social media data requires comprehensive cleaning to manage informal language, emojis, and abbreviations effectively. The relationship between pre-processing rigor and improved outcomes is emphasized that can lead to higher accuracy and F1-scores, it can also increase computational cost and complexity.

Preprocessing is an essential step in preparing Twitter data for effective sentiment analysis due to its noisy and unstructured nature. Tokenization, which breaks text into smaller units, is critical for managing hashtags, mentions, and punctuation effectively. Handling emojis, often significant carriers of sentiment, involves converting them into textual equivalents or sentiment scores. Hashtags, which encapsulate key sentiments or topics, require splitting into component words for accurate analysis (e.g., ''HappyDay'' becomes ''Happy Day''). Noise removal, including eliminating retweets, URLs, and irrelevant symbols, improves data quality and reduces distractions for models. Normalization processes, such as standardizing abbreviations, lowercase text, and handling elongated words (e.g. ''cooool'' ''cool''), ensure uniformity in the dataset. These preprocessing techniques enhance the interpretability and performance of NLP models applied to Twitter sentiment analysis.

C. IMPLEMENTATION PRACTICES, HYPER-PARAMETERS, AND DATASETS IN REVIEWED STUDIES

As this paper is a comprehensive review, we have synthesized the implementation processes, hyper-parameters, and datasets commonly utilized in sentiment analysis studies to provide actionable insights and context for researchers. These aspects are integral to understanding the performance and applicability of various natural language processing (NLP) methods for sentiment analysis tasks, particularly on Twitter data.

D. DATASETS USED IN REVIEWED STUDIES

A variety of datasets have been employed in the studies we reviewed, each catering to different sentiment analysis objectives and language-specific tasks. Among the most popular datasets is **Sentiment140**, a benchmark dataset that contains 1.6 million labeled tweets (positive and negative), making it a cornerstone for Twitter-specific sentiment analysis. The dataset is frequently used to evaluate traditional machine learning methods as well as modern deep learning architectures. Another widely used dataset is the **SemEval series**, which includes domain-specific tasks and multilingual sentiment datasets, providing a robust platform for evaluating the adaptability of models across languages and contexts. For general sentiment analysis, datasets like the **IMDB Movie Reviews** (50,000 labeled reviews) and **Amazon Product Reviews** (millions of customer reviews with sentiment labels) are employed to benchmark models on longer-form text.

Other datasets focus on specific domains, such as **COVID-19 Twitter datasets** that analyze public sentiment during crises, and the **Reddit Comment Corpus**, which enables sentiment analysis in informal, user-generated long-form text. Each of these datasets presents unique challenges, including class imbalance, informal language, and mixed sentiments, requiring advanced techniques to achieve accurate classification.

E. PREPROCESSING TECHNIQUES

Preprocessing plays a critical role in preparing textual data, especially for Twitter sentiment analysis, where the language is often noisy and informal. Most studies we reviewed employ a series of preprocessing steps, including:

- **Noise Removal:** Elimination of URLs, mentions (e.g., `@username`), and retweets to reduce irrelevant features.
- **Tokenization:** Splitting text into smaller units, such as words or subwords, using methods like WordPiece or Byte Pair Encoding (BPE).
- **Stopword Removal:** Filtering out common words (e.g., `and`, `the`) that do not carry significant sentiment information.
- **Emoji and Hashtag Handling:** Converting emojis into textual equivalents (e.g., → `happy`) and splitting hashtags into component words (e.g., `#HappyDay` → `Happy Day`).
- **Normalization:** Lowercasing text, standardizing abbreviations, and handling elongated words (e.g., `cooool` → `cool`).
- **Data Balancing:** Techniques such as oversampling, undersampling, or synthetic data generation are used to address class imbalance issues.

These preprocessing steps significantly improve model performance by reducing noise and standardizing input data. In some cases, advanced techniques such as **back-translation** and **synonym replacement** are employed for data augmentation, enhancing model robustness.

F. IMPLEMENTATION AND HYPER-PARAMETERS IN REVIEWED STUDIES

The reviewed studies highlight a range of implementation practices and hyper-parameter configurations that are critical for training effective sentiment analysis models. For state-of-the-art transformer-based architectures, such as **BERT** and **RoBERTa**, fine-tuning is typically performed with the following hyper-parameters:

- **Learning Rate:** Most studies use a small learning rate in the range of 2×10^{-5} to 5×10^{-5}, which prevents overfitting during fine-tuning on domain-specific datasets.

- **Batch Size:** Common batch sizes include 16 or 32, balancing memory constraints and training efficiency.
- **Epochs:** Fine-tuning is often conducted over 3 to 5 epochs, as longer training can lead to overfitting on small datasets.
- **Optimizer:** The **Adam** optimizer, particularly its variant **AdamW**, is frequently used due to its adaptive learning rate and regularization capabilities.
- **Max Sequence Length:** Input sequences are typically truncated or padded to 128 or 256 tokens to fit within memory constraints while preserving sufficient context for sentiment classification.

For models like **GPT** and **GPT-3.5**, hyper-parameters include larger context windows (e.g., 512 or 1024 tokens) and specialized learning rate schedules. Studies leveraging ensemble methods, such as combining **BERT** with **LSTM** or **CNN**, often focus on optimizing hyper-parameters for both components to maximize complementary strengths.

G. EVALUATION METRICS

Evaluation is a key aspect of sentiment analysis studies, and the reviewed literature commonly reports the following metrics:

- **Accuracy:** A widely used metric for balanced datasets, indicating the proportion of correctly classified instances.
- **F1-Score:** Particularly important for imbalanced datasets, combining precision and recall into a single metric.
- **Cross-Entropy Loss:** Used during training to quantify the difference between predicted and true class probabilities.
- **Area Under the Curve (AUC):** Evaluates the model's ability to distinguish between classes across different thresholds.

H. INSIGHTS AND IMPLICATIONS

The reviewed studies demonstrate that achieving high performance in sentiment analysis requires careful attention to dataset selection, preprocessing, and hyper-parameter tuning. Models like **BERT**, **RoBERTa**, and **GPT-3.5** consistently outperform traditional methods, especially when fine-tuned on domain-specific or multilingual datasets. However, challenges such as real-time analysis, handling noisy data, and computational efficiency remain active areas of research. This synthesis of implementation practices aims to guide future studies and provide researchers with actionable insights to design effective sentiment analysis pipelines.

I. CHALLENGES

Despite advances in NLP model capabilities, several challenges persist in sentiment analysis. One of the main issues discussed is the difficulty in detecting nuanced expressions such as sarcasm, irony, and mixed sentiments, which often require a level of language understanding that current models struggle to achieve. Although transformer-based models have advanced context understanding, they are not infallible when external knowledge or cultural context is necessary for accurate interpretation. Another challenge is the limitation in cross-domain performance; models trained on a specific dataset often perform suboptimally when applied to different domains, which limits their generalizability and scalability. This section also addresses computational and resource constraints, particularly for large models like GPT-3, which require significant processing power and data to fine-tune effectively. Finally, the discussion highlights ethical concerns, including biases in training data that can influence model output and reinforce harmful stereotypes.

J. MODEL GENERALIZATION AND ADAPTABILITY

Generalization across different datasets and domains is a key measure of the robustness of an NLP model. This subsection reviews how models like BERT and RoBERTa have demonstrated adaptability in various tasks but may require domain-specific fine-tuning to maintain performance when applied to new data types. It explores strategies for enhancing cross-domain generalization, such as transfer learning, domain adaptation, and using ensemble models that combine the strengths of multiple architectures. Additionally, it discusses the importance of creating more diverse and representative training datasets to improve model adaptability. The discussion emphasizes that while models like SBERT and BiLSTM-RNN hybrids have shown promise in balancing generalization with performance, further research is needed to develop models that can consistently perform well across different sentiment analysis scenarios.

K. ETHICAL CONSIDERATIONS AND BIAS IN MODEL TRAINING

Ethical issues related to the training of the NLP model are crucial, as biases in training data can lead to biased outcomes that reinforce social stereotypes or disadvantage certain groups. This section explores the sources of such biases, which may stem from unbalanced data sets or inherent biases in user-generated content. The impact of these biases on sentiment analysis can result in skewed classifications, particularly when analyzing sentiments related to sensitive topics. The discussion advocates for the integration of bias detection tools and fairness benchmarks as part of the model development and evaluation process. Techniques such as adversarial training and data augmentation strategies aimed at reducing biases are examined, along with calls for transparency in dataset curation and algorithm development. Addressing these ethical concerns is essential to build trust in the NLP systems used in sentiment analysis and to ensure equitable outcomes across different demographic groups of users.

L. PRACTICAL IMPLICATIONS FOR INDUSTRY AND RESEARCH

The practical applications of sentiment analysis using advanced NLP models extend across various industries, including marketing, customer service, politics, and public health. In marketing, companies leverage sentiment analysis to monitor brand perception and respond to customer feedback in real time, helping them refine their strategies and improve customer satisfaction. In public health and policymaking, governments and organizations can use sentiment analysis to track public opinion on initiatives, assess community concerns, and make informed decisions. This section represent how businesses and researchers can apply the insights from this review to select appropriate models, refine pre-processing pipelines, and adapt their approaches to specific use cases. It also discusses the implications of using models at scale, including the need for robust infrastructure and ongoing evaluation to ensure reliable outputs.

M. SUMMARY OF KEY FINDINGS

The review highlights that transformer-based models, such as BERT and RoBERTa, outperform traditional approaches due to their ability to capture deep contextual relationships. BERT's bidirectional training allows it to understand complex sentence structures, making it highly effective for nuanced sentiment tasks, while RoBERTa achieves superior performance in multiclass sentiment classification with enhanced training methods. GPT variants excel in handling nuanced context and generating human-like responses, but require extensive fine-tuning for optimal performance on Twitter data. Traditional models like LSTM and CNN are effective for sequential data processing, but fall short in capturing deep context compared to transformer architectures. Pre-processing steps, including tokenization and handling hashtags or emojis, directly impact model accuracy and robustness. However, challenges such as understanding sarcasm, managing domain shifts, and addressing computational resource demands highlight areas for future research and development.

V. CONCLUSION AND FUTURE WORKS

In this paper, we conducted a comprehensive review of sentiment analysis using advanced NLP models, including BERT, GPT variants, RoBERTa, and hybrid approaches. Our analysis covered various aspects of these models, such as their application to different datasets, pre-processing techniques, performance metrics, and key findings from recent studies. While transformer-based models like BERT and RoBERTa demonstrate strong capabilities in handling complex linguistic patterns and achieving high performance in sentiment classification tasks, their effectiveness is significantly influenced by the nature of the dataset, pre-processing methods, and the domain of application.

Transformer-based architectures, particularly BERT and RoBERTa, have shown substantial promise in surpassing traditional machine learning models and simpler deep learning frameworks, thanks to their deep contextual understanding and bidirectional training. Models such as GPT-3 and its successors have also displayed significant potential in handling context-rich text and generating human-like content. However, these models often require extensive fine-tuning to achieve optimal results in sentiment analysis, underscoring the importance of tailored pre-processing techniques and domain-specific adaptation. Despite the advancements, challenges such as understanding nuanced language constructs like sarcasm, irony, and mixed sentiments, as well as handling bias and ethical concerns, remain prominent. The variability in cross-domain performance highlights the need for more adaptive and generalizable approaches.

To advance the field of sentiment analysis, future research should focus on several key areas:

Improving Model Generalization and Cross-Domain Performance: Research should explore hybrid and ensemble approaches that combine the strengths of different models to create more adaptable solutions capable of maintaining high performance across various datasets and domains. The development of transfer learning techniques and cross-lingual training frameworks can further enhance model adaptability.

Advanced Pre-Processing Techniques: Future studies should aim to develop more sophisticated pre-processing pipelines that are capable of managing informal language, slang, emojis, and context-specific data prevalent in social media and other user-generated content. Techniques that incorporate external knowledge bases and context-aware data cleaning strategies can greatly improve model outputs.

Incorporation of Multimodal Data: Integrating textual data with other modalities, such as images, audio, or video, can provide richer context and improve the ability of models to interpret sentiments accurately. Multimodal models can capture additional nuances and offer a more comprehensive understanding of user sentiment, particularly in platforms where text is accompanied by visual or auditory cues.

Ethical Considerations and Bias Mitigation: Addressing biases inherent in training data and ensuring that models operate fairly across different user demographics is crucial for building trust in NLP-based sentiment analysis tools. Future work should emphasize the integration of ethical evaluation frameworks and bias detection mechanisms during model development and training.

Real-Time and Scalable Solutions: The implementation of NLP models for large-scale, real-time sentiment analysis requires efficient and scalable solutions. Future research should explore lightweight models and optimization techniques that reduce the computational overhead while maintaining performance. This is particularly relevant for industries that require rapid sentiment tracking to inform decision-making.

Explainability and Interpretability: As the complexity of NLP models increases, so does the importance of understanding how they arrive at specific outputs. Future works should focus on enhancing model interpretability,

providing insights into which features contribute most to sentiment predictions. This can help users and stakeholders trust and validate the decisions made by these systems.

The evolving landscape of NLP continues to open new possibilities for sentiment analysis, with transformer-based architectures leading the way in innovation. However, to fully harness the power of these models, it is crucial to address existing challenges, develop more adaptive and context-aware solutions, and incorporate robust ethical standards. Future research and development in these areas will help pave the way for more accurate, fair, and practical sentiment analysis applications, extending their impact across industries and research domains. The continued evolution of techniques, along with collaborative efforts between researchers and practitioners, will contribute to more comprehensive, reliable, and equitable sentiment analysis solutions.

REFERENCES

[1] D. A. Gruber, R. E. Smerek, M. C. Thomas-Hunt, and E. H. James, "The real-time power of Twitter: Crisis management and leadership in an age of social media," *Bus. Horizons*, vol. 58, no. 2, pp. 163–172, Mar. 2015.

[2] P. Pond and J. Lewis, "Riots and Twitter: Connective politics, social media and framing discourses in the digital public sphere," *Inf., Commun. Soc.*, vol. 22, no. 2, pp. 213–231, Jan. 2019.

[3] M. Martínez-Rojas, M. D. C. Pardo-Ferreira, and J. C. Rubio-Romero, "Twitter as a tool for the management and analysis of emergency situations: A systematic literature review," *Int. J. Inf. Manage.*, vol. 43, pp. 196–208, Dec. 2018.

[4] L. Potts and S. Mahnke, "Subverting the platform flexibility of Twitter to spread misinformation," in *Platforms, Protests, and the Challenge of Networked Democracy*, 2020, pp. 157–172.

[5] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," 2009, *arXiv:0911.1583*.

[6] B. Balducci and D. Marinova, "Unstructured data in marketing," *J. Acad. Marketing Sci.*, vol. 46, no. 4, pp. 557–590, Jul. 2018.

[7] J. Hurlock and M. Wilson, "Searching Twitter: Separating the tweet from the chaff," in *Proc. Int. AAAI Conf. Web Social Media*, Aug. 2021, vol. 5, no. 1, pp. 161–168.

[8] P. K. Jain, V. Saravanan, and R. Pamula, "A hybrid CNN-LSTM: A deep learning approach for consumer sentiment analysis using qualitative user-generated contents," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–15, Sep. 2021.

[9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.

[10] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[11] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," *Inf. Fusion*, vol. 44, pp. 65–77, Nov. 2018.

[12] M. Sykora, S. Elayan, I. R. Hodgkinson, T. W. Jackson, and A. West, "The power of emotions: Leveraging user generated content for customer experience management," *J. Bus. Res.*, vol. 144, pp. 997–1006, May 2022.

[13] S. Mishra, S. Choubey, A. Choubey, N. Yogeesh, J. D. P. Rao, and P. William, "Data extraction approach using natural language processing for sentiment analysis," in *Proc. Int. Conf. Autom., Comput. Renew. Syst. (ICACRS)*, Dec. 2022, pp. 970–972.

[14] S. Scheidt and Q. B. Chung, "Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation," *Int. J. Inf. Manage.*, vol. 45, pp. 223–232, Apr. 2019.

[15] J. Manurung, M. H. Napitupulu, and H. Simangunsong, "Exploring the impact of slang usage among students on WhatsApp: A dig-ital linguistic analysis," *Jurnal Ilmu Pendidikan dan Humaniora*, vol. 11, no. 2, pp. 153–169, May 2022.

[16] M. Zappavigna, *Searchable Talk: Hashtags and Social Media Metadiscourse*. Bloomsbury, 2018.

[17] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 845–863, Apr. 2022.

[18] M. Sykora, S. Elayan, and T. W. Jackson, "A qualitative analysis of sarcasm, irony and related #hashtags on Twitter," *Big Data Soc.*, vol. 7, no. 2, Jul. 2020, Art. no. 2053951720972735.

[19] L. Weitzel, R. C. Prati, and R. F. Aguiar, "The comprehension of figurative language: What is the influence of irony and sarcasm on NLP techniques?" in *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence*, 2016, pp. 49–74.

[20] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Lang. Process. J.*, vol. 4, Sep. 2023, Art. no. 100026.

[21] Y. Shu, Y. Ma, W. Li, G. Hu, X. Wang, and Q. Zhang, "Unraveling the dynamics of social governance innovation: A synergistic approach employing NLP and network analysis," *Exp. Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124632.

[22] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl. Inf. Syst.*, vol. 60, no. 2, pp. 617–663, Jul. 2018.

[23] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 2, pp. 450–464, Apr. 2020.

[24] A. K. Rathore, A. K. Kar, and P. V. Ilavarasan, "Social media analytics: Literature review and directions for future research," *Decis. Anal.*, vol. 14, no. 4, pp. 229–249, Dec. 2017.

[25] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, *Sentiment Analysis in Social Networks*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[26] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. 2nd Int. Conf. Knowl. Capture*, 2003, pp. 70–77.

[27] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can large language models transform computational social science?" *Comput. Linguistics*, vol. 50, no. 1, pp. 237–291, Mar. 2024.

[28] R. M. Devadas, V. Hiremani, J. P. Gujjar, N. S. Rani, and K. Bhavya, "Innovative fusion: Attention-augmented support vector machines for superior text classification for social marketing," in *Advances in Data Analytics for Influencer Marketing: An Interdisciplinary Approach*. Berlin, Germany: Springer, 2024, pp. 283–303.

[29] G. I. Ahmad, J. Singla, A. Ali, A. A. Reshi, and A. A. Salameh, "Machine learning techniques for sentiment analysis of code-mixed and switched Indian social media text corpus—A comprehensive review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, 2022.

[30] N. Chawla and V. Vansh, "Unveiling emotions through sentiment analysis," Tech. Rep., 2024.

[31] A. Hassan and A. Mahmood, "Convolutional recurrent deep learning model for sentence classification," *IEEE Access*, vol. 6, pp. 13949–13957, 2018.

[32] E. Omara, M. Mousa, and N. Ismail, "Character gated recurrent neural networks for Arabic sentiment analysis," *Sci. Rep.*, vol. 12, no. 1, p. 9779, Jun. 2022.

[33] Y. Liu, L. Wang, T. Shi, and J. Li, "Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and bi-LSTM," *Inf. Syst.*, vol. 103, Jan. 2022, Art. no. 101865.

[34] S. Gajendran, D. Manjula, and V. Sugumaran, "Character level and word level embedding with bidirectional LSTM—Dynamic recurrent neural network for biomedical named entity recognition from literature," *J. Biomed. Informat.*, vol. 112, Dec. 2020, Art. no. 103609.

[35] I. Kondurkar, A. Raj, and D. Lakshmi, "Modern applications with a focus on training ChatGPT and GPT models: Exploring generative AI and NLP," in *Advanced Applications of Generative AI and Natural Language Processing Models*. Hershey, PA, USA: IGI Global, 2024, pp. 186–227.

[36] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," 2023, *arXiv:2302.09419*.

[37] K. Machová, I. Srba, M. Sarnovský, J. Paralič, V. Maslej-Kreşňáková, A. Hrč ková, M. Kompan, M. Šimko, R. Blaho, D. Chudá, M. Bieliková, and P. Návrat, "Addressing false information and abusive language in digital space using intelligent approaches," in *Towards Digital Intelligence Society: A Knowledge-Based Approach*. Berlin, Germany: Springer, 2021, pp. 3–32.

[38] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on Twitter," *Multimedia Tools Appl.*, vol. 80, nos. 28–29, pp. 35239–35266, Nov. 2021.

[39] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: A review," *Artif. Intell. Rev.*, vol. 56, no. 9, pp. 10345–10425, Sep. 2023.

[40] J. Choi, J. Yoon, J. Chung, B.-Y. Coh, and J.-M. Lee, "Social media analytics and business intelligence research: A systematic review," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102279.

[41] E. L. Jenkins, D. Lukose, L. Brennan, A. Molenaar, and T. A. McCaffrey, "Exploring food waste conversations on social media: A sentiment, emotion, and topic analysis of Twitter data," *Sustainability*, vol. 15, no. 18, p. 13788, Sep. 2023.

[42] S. Cartwright, H. Liu, and C. Raddats, "Strategic use of social media within business-to-business (B2B) marketing: A systematic literature review," *Ind. Marketing Manage.*, vol. 97, pp. 35–58, Aug. 2021.

[43] J. Hruska and P. Maresova, "Use of social media platforms among adults in the United States—Behavior on social media," *Societies*, vol. 10, no. 1, p. 27, Mar. 2020.

[44] A. Castillo, J. Benitez, J. Liorens, and J. Braojos, "Impact of social media on the firm's knowledge exploration and knowledge exploitation: The role of business analytics talent," *J. Assoc. Inf. Syst.*, vol. 22, no. 5, pp. 1472–1508, 2021.

[45] A. Kumar and A. Jaiswal, "Systematic literature review of sentiment analysis on Twitter using soft computing techniques," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 1, p. 5107, Jan. 2020.

[46] J. Hartmann, M. Heitmann, C. SieBERT, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," *Int. J. Res. Marketing*, vol. 40, no. 1, pp. 75–87, Mar. 2023.

[47] A. Goel, J. Gautam, and S. Kumar, "Real time sentiment analysis of tweets using naive Bayes," in *Proc. 2nd Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Oct. 2016, pp. 257–261.

[48] B. Zeng and R. Gerritsen, "What do we know about social media in tourism? A review," *Tourism Manage. Perspect.*, vol. 10, pp. 27–36, Apr. 2014.

[49] P. Mehta and S. Pandya, "A review on sentiment analysis methodologies, practices and applications," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 601–609, Feb. 2020.

[50] S. Steinert, "Corona and value change. The role of social media and emotional contagion," *Ethics Inf. Technol.*, vol. 23, no. 1, pp. 59–68, Nov. 2021.

[51] G. A. V. Kleef and S. Cŏté, "The social effects of emotions," *Annu. Rev. Psychol.*, vol. 73, no. 1, pp. 629–658, Jul. 2021.

[52] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, "Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models," *Eur. J. Med. Res.*, vol. 29, no. 1, p. 455, Sep. 2024.

[53] Z. A. Saberi, H. Sadr, and M. R. Yamaghani, "An intelligent diagnosis system for predicting coronary heart disease," in *Proc. 10th Int. Conf. Artif. Intell. Robot. (QICAR)*, Feb. 2024, pp. 131–137.

[54] Z. Khodaverdian, H. Sadr, and S. A. Edalatpanah, "A shallow deep neural network for selection of migration candidate virtual machines to reduce energy consumption," in *Proc. 7th Int. Conf. Web Res. (ICWR)*, May 2021, pp. 191–196.

[55] M. Nazari, H. Emami, R. Rabiei, A. Hosseini, and S. Rahmatizadeh, "Detection of cardiovascular diseases using data mining approaches: Application of an ensemble-based model," *Cognit. Comput.*, vol. 16, no. 5, pp. 2264–2278, Sep. 2024.

[56] M. Nazari, S. Moayed Rezaie, F. Yaseri, H. Sadr, and E. Nazari, "Design and analysis of a telemonitoring system for high-risk pregnant women in need of special care or attention," *BMC Pregnancy Childbirth*, vol. 24, no. 1, p. 817, Dec. 2024.

[57] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of naive Bayes and SVM algorithm based on sentiment analysis using review dataset," in *Proc. 8th Int. Conf. Syst. Model. Advancement Res. Trends (SMART)*, Nov. 2019, pp. 266–270.

[58] M. M. Alnaddaf and M. S. Başarslan, "Sentiment analysis using various machine learning techniques on depression review data," in *Proc. 8th Int. Artif. Intell. Data Process. Symp. (IDAP)*, vol. 4, Sep. 2024, pp. 1–5.

[59] H. Shrestha, C. Dhasarathan, S. Munisamy, and A. Jayavel, "Natural language processing based sentimental analysis of Hindi (SAH) script an optimization approach," *Int. J. Speech Technol.*, vol. 23, no. 4, pp. 757–766, Dec. 2020.

[60] Z. Hu, I. Dychka, K. Potapova, and V. Meliukh, "Augmenting sentiment analysis prediction in binary text classification through advanced natural language processing models and classifiers," *Int. J. Inf. Technol. Comput. Sci.*, vol. 16, no. 2, pp. 16–31, Apr. 2024.

[61] J. Jia, W. Liang, and Y. Liang, "A review of hybrid and ensemble in deep learning for natural language processing," 2023, *arXiv:2312.05589*.

[62] M. Hajiali, "Big data and sentiment analysis: A comprehensive and systematic literature review," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 14, p. 5671, Jul. 2020.

[63] W. Ansar, S. Goswami, and A. Chakrabarti, "A survey on transformers in NLP with focus on efficiency," 2024, *arXiv:2406.16893*.

[64] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea Preprints*, Dec. 2024.

[65] B. Desai, K. Patil, A. Patil, and I. Mehta, "Large language models: A comprehensive exploration of modern AI's potential and pitfalls," *J. Innov. Technol.*, vol. 6, no. 1, 2023.

[66] F. Shamrat, S. Chakraborty, M. Imran, J. N. Muna, M. M. Billah, P. Das, and O. Rahman, "Sentiment analysis on Twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 23, no. 1, pp. 463–470, 2021.

[67] J.-W. Chang, N. Yen, and J. C. Hung, "Design of a NLP-empowered finance fraud awareness model: The anti-fraud chatbot for fraud detection and fraud classification as an instance," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 10, pp. 4663–4679, Oct. 2022.

[68] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Comput. Surv.*, vol. 56, no. 2, pp. 1–40, Feb. 2024.

[69] N. Sharma and B. Verma, "Recent advances in transfer learning for natural language processing (NLP)," in *Federated Learning for Internet of Vehicles: IoV Image Processing, Vision and Intelligent Systems*, 2024, pp. 228–254.

[70] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020.

[71] F.-E. Lagrari and Y. ElKettani, "A comparative study of a new customized BERT for sentiment analysis," in *Sentiment Analysis and Deep Learning*. Berlin, Germany: Springer, 2023, pp. 315–322.

[72] S. M. Qaisar, "Sentiment analysis of IMDb movie reviews using long short-term memory," in *Proc. 2nd Int. Conf. Comput. Inf. Sci. (ICCIS)*, Oct. 2020, pp. 1–4.

[73] M. A. Jahin, M. S. H. Shovon, M. F. Mridha, M. R. Islam, and Y. Watanobe, "A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets," 2024, *arXiv:2404.00297*.

[74] S. Kashid, K. Kumar, P. Saini, A. Dhiman, and A. Negi, "Bi-RNN and bi-LSTM based text classification for Amazon reviews," in *Proc. Int. Conf. Deep Learn., Artif. Intell. Robot.* Springer, 2022, pp. 62–72.

[75] S. Alipour, A. Galeazzi, E. Sangiorgio, M. Avalle, L. Bojic, M. Cinelli, and W. Quattrociocchi, "Cross-platform social dynamics: An analysis of ChatGPT and COVID-19 vaccine conversations," *Sci. Rep.*, vol. 14, no. 1, p. 2789, Feb. 2024.

[76] A. Amini, Y. E. Bayiz, A. Ram, R. Marculescu, and U. Topcu, "News source credibility assessment: A Reddit case study," 2024, *arXiv:2402.10938*.

[77] A. Sittar, D. Mladenić, and M. Grobelnik, "Profiling the barriers to the spreading of news using news headlines," *Frontiers Artif. Intell.*, vol. 6, Aug. 2023, Art. no. 1225213.

[78] M. Qorich and R. El Ouazzani, "Text sentiment classification of Amazon reviews using word embeddings and convolutional neural networks," *J. Supercomput.*, vol. 79, no. 10, pp. 11029–11054, Jul. 2023.

[79] F. Nadi, H. Naghavipour, T. Mehmood, A. B. Azman, J. A. P. Nagantheran, K. S. K. Ting, N. M. I. B. N. Adnan, R. A. P. Sivarajan, S. A. P. Veerah, and R. F. Rahmat, "Sentiment analysis using large language models: A case study of GPT-3.5," in *Proc. Int. Conf. Data Sci. Emerg. Technol.* Springer, 2024, pp. 161–168.

[80] J. Jiang, X. Ren, and E. Ferrara, "Retweet-BERT: Political leaning detection using language features and information diffusion on social networks," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 17, Jun. 2023, pp. 459–469.

[81] J.-C. Na, W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, and Y.-L. Theng, "Sentiment classification of drug reviews using a rule-based linguistic approach," in *Proc. 14th Int. Conf. Asia–Pacific Digital Libraries*, Taipei, Taiwan. Springer, Nov. 2012, pp. 189–198.

[82] T. K. Sonali Ridhorkar, "RMDEASD: Integrating rule mining and deep learning for enhanced aspect-based sentiment analysis across diverse domains," *J. Electr. Syst.*, vol. 20, no. 3, pp. 1163–1192, Apr. 2024.

[83] S. de la Harpe, R. Palermo, E. Brown, N. Fay, and A. Dawel, "People attribute a range of highly-varied and socially-bound meanings to naturalistic sad facial expressions," *J. Nonverbal Behav.*, vol. 48, no. 3, pp. 465–493, Sep. 2024.

[84] P. Berka, "Sentiment analysis using rule-based and case-based reasoning," *J. Intell. Inf. Syst.*, vol. 55, no. 1, pp. 51–66, Aug. 2020.

[85] H. Rahab, H. Haouassi, and A. Laouid, "Rule-based Arabic sentiment analysis using binary equilibrium optimization algorithm," *Arabian J. Sci. Eng.*, vol. 48, no. 2, pp. 2359–2374, Feb. 2023.

[86] R. Saha, O. Granmo, and M. Goodwin, "Mining interpretable rules for sentiment and semantic relation analysis using tsetlin machines," in *Proc. Int. Conf. Innov. Techn. Appl. Artif. Intell.* Springer, 2020, pp. 67–78.

[87] Z. Zheng, Y.-C. Zhou, K.-Y. Chen, X.-Z. Lu, Z.-T. She, and J.-R. Lin, "A text classification-based approach for evaluating and enhancing the machine interpretability of building codes," *Eng. Appl. Artif. Intell.*, vol. 127, Jan. 2024, Art. no. 107207.

[88] P. Monika, C. Kulkarni, N. H. Kumar, S. Shruthi, and V. Vani, "Machine learning approaches for sentiment analysis: A survey," *Int. J. Health Sci.*, vol. 6, no. S4, pp. 1286–1300, 2022.

[89] D. M. Abdullah and A. M. Abdulazeez, "Machine learning applications based on SVM classification a review," *Qubahan Academic J.*, vol. 1, no. 2, pp. 81–90, Apr. 2021.

[90] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," *Information*, vol. 15, no. 9, p. 517, Aug. 2024.

[91] N. E. Michael, R. C. Bansal, A. A. A. Ismail, A. Elnady, and S. Hasan, "A cohesive structure of bi-directional long-short-term memory (BiLSTM)—GRU for predicting hourly solar radiation," *Renew. Energy*, vol. 222, Feb. 2024, Art. no. 119943.

[92] M. Jiang, J. Wu, X. Shi, and M. Zhang, "Transformer based memory network for sentiment analysis of web comments," *IEEE Access*, vol. 7, pp. 179942–179953, 2019.

[93] H. Holm, "Bidirectional encoder representations from transformers (BERT) for question answering in the telecom domain: Adapting a BERT-like language model to the telecom domain using the ELECTRA pre-training approach," Tech. Rep., 2021.

[94] H. Shim, D. Lowet, S. Luca, and B. Vanrumste, "LETS: A label-efficient training scheme for aspect-based sentiment analysis by using a pre-trained language model," *IEEE Access*, vol. 9, pp. 115563–115578, 2021.

[95] A. Kalinowski, "Developing novel triple embeddings for scalable alignment of knowledge graphs and natural language," Ph.D. dissertation, Drexel Univ., 2024.

[96] S. S. Sundaram, S. Gurajada, D. Padmanabhan, S. S. Abraham, and M. Fisichella, "Does a language model 'understand' high school math? A survey of deep learning based word problem solvers," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, p. e1534, Mar. 2024.

[97] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and P. Group, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Int. J. Surg.*, vol. 8, no. 5, pp. 336–341, 2010.

[98] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, P. Badhani, and B. Tech, "Study of Twitter sentiment analysis using machine learning algorithms on Python," *Int. J. Comput. Appl.*, vol. 165, no. 9, pp. 29–34, 2017.

[99] R. B. Saranya, R. Kesavan, and K. N. Devi, "Extremely randomized tree based sentiment polarity classification on online product reviews," in *Proc. Int. Conf. Big Data Analytics*. Springer, 2022, pp. 159–171.

[100] D. Hazarika, G. Konwar, S. Deb, and D. J. Bora, "Sentiment analysis on Twitter by using TextBlob for natural language processing," *ICRMAT*, vol. 24, pp. 63–67, Jan. 2020.

[101] S. Rathje, D.-M. Mirea, I. Sucholutsky, R. Marjieh, C. E. RoBERTson, and J. J. Van Bavel, "GPT is an effective tool for multilingual psychological text analysis," *Proc. Nat. Acad. Sci. USA*, vol. 121, no. 34, Aug. 2024, Art. no. 2308950121.

[102] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning," 2023, *arXiv:2307.10234*.

[103] G. I. Ahmad, J. Singla, and N. Nikita, "Review on sentiment analysis of Indian languages with a special focus on code mixed Indian languages," in *Proc. Int. Conf. Autom., Comput. Technol. Manage. (ICACTM)*, Apr. 2019, pp. 352–356.

[104] M. E. Chatzimina, H. A. Papadaki, C. Pontikoglou, and M. Tsiknakis, "A comparative sentiment analysis of Greek clinical conversations using BERT, RoBERTa, GPT-2, and XLNet," *Bioengineering*, vol. 11, no. 6, p. 521, May 2024.

[105] R. Chandra and A. Krishna, "COVID-19 sentiment analysis via deep learning during the rise of novel cases," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0255615.

[106] I. El Karfi and S. El Fkihi, "A combined Bi-LSTM-GPT model for Arabic sentiment analysis," *Int. J. Intell. Syst. Appl. Eng.*, vol. 11, pp. 77–84, Jan. 2023.

[107] M. Pota, M. Ventura, R. Catelli, and M. Esposito, "An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian," *Sensors*, vol. 21, no. 1, p. 133, Dec. 2020.

[108] S. Ahmed, M. M. Samia, M. H. Sayma, M. M. Kabir, and M. F. Mridha, "TRF-BERT: A transformative approach to aspect-based sentiment analysis in the Bengali language," *PLoS ONE*, vol. 19, no. 9, Sep. 2024, Art. no. e0308050.

[109] K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," *Kurdistan J. Appl. Res.*, pp. 54–65, May 2020.

[110] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter," *Frontiers Artif. Intell.*, vol. 6, Mar. 2023, Art. no. 1023281.

[111] M. Heidari and J. H. Jones, "Using BERT to extract topic-independent sentiment features for social media bot detection," in *Proc. 11th IEEE Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2020, pp. 542–547.

[112] H. Meisheri and L. Dey, "TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 291–299.

[113] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019.

[114] Venkatesh, S. U. Hegde, A. S. Zaiba, and Y. Nagaraju, "Hybrid CNN-LSTM model with GloVe word vector for sentiment analysis on football specific tweets," in *Proc. Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Feb. 2021, pp. 1–8.

[115] A. Karimi, L. Rossi, and A. Prati, "Adversarial training for aspect-based sentiment analysis with BERT," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8797–8803.

[116] L. Zhang, H. Fan, C. Peng, G. Rao, and Q. Cong, "Sentiment analysis methods for HPV vaccines related tweets based on transfer learning," in *Proc. MDPI*, Aug. 2020, vol. 8, no. 3, p. 307.

[117] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using BERT," in *Proc. EDBT/ICDT Workshops*, Jan. 2021, pp. 1–7.

[118] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.

[119] U. Sirisha and B. S. Chandana, "Aspect based sentiment & emotion analysis with ROBERTa, LSTM," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 11, 2022.

[120] M. A. Jahin, M. S. H. Shovon, and M. F. Mridha, "TRABSA: Interpretable sentiment analysis of tweets using attention-based BiLSTM and Twitter-RoBERTa," 2024, *arXiv:2404.00297*.

[121] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A hybrid deep learning model for enhanced sentiment analysis," *Appl. Sci.*, vol. 13, no. 6, p. 3915, Mar. 2023.

[122] C.-H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *J. Big Data*, vol. 10, no. 1, p. 88, May 2023.

[123] M. L. Jamil, S. Pais, J. Cordeiro, and G. Dias, "Detect extreme sentiments on social networks using BERT," Tech. Rep., 2021.

[124] S. S. Ayon, S. Ishrat, S. A. Mallick, P. C. Das, and F. B. Ashraf, "Sentiment analysis on COVID-19 tweets," in *Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2022, pp. 551–556.

[125] M. Rao, A. Kumar, and V. Tyagi, "Sentiment analysis of user-generated data using CNN-BiLSTM model," in *Proc. Int. Conf. Adv. Commun. Intell. Syst.* Springer, 2023, pp. 239–246.

[126] A. J. Keya, H. H. Shajeeb, M. S. Rahman, and M. F. Mridha, "FakeStack: Hierarchical tri-BERT-CNN-LSTM stacked model for effective fake news detection," *PLoS ONE*, vol. 18, no. 12, Dec. 2023, Art. no. e0294701.

[127] S. K. Nair and R. Soni, "Sentiment analysis on movie reviews using recurrent neural network," *IRE J.*, vol. 1, no. 10, 2018.

[128] G. Kumar, R. Agrawal, K. Sharma, P. R. Gundalwar, A. Kazi, P. Agrawal, M. Tomar, and S. Salagrama, "Combining BERT and CNN for sentiment analysis a case study on COVID-19," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 10, 2024.

[129] A. Aiswarya and H. Rajeev, "YouTube comment sentimental analysis," *Indian J. Data Mining*, vol. 4, no. 1, pp. 5–8, 2024.

[130] A. Areshey and H. Mathkour, "Transfer learning for sentiment classification using bidirectional encoder representations from transformers (BERT) model," *Sensors*, vol. 23, no. 11, p. 5232, May 2023.

[131] J. O. Krugmann and J. Hartmann, "Sentiment analysis in the age of generative AI," *Customer Needs Solutions*, vol. 11, no. 1, p. 3, Dec. 2024.

[132] Z. Liu, "Yelp review rating prediction: Machine learning and deep learning models," 2020, *arXiv:2012.06690*.

[133] A. R. Abas, I. Elhenawy, M. Zidan, and M. Othman, "BERT-CNN: A deep learning model for detecting emotions from text," *Comput., Mater. Continua*, vol. 71, no. 2, pp. 2943–2961, 2022.

[134] D. Rozado, R. Hughes, and J. Halberstadt, "Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models," *PLoS ONE*, vol. 17, no. 10, Oct. 2022, Art. no. e0276367.

[135] Z. Chen, R. Yang, S. Fu, N. Zong, H. Liu, and M. Huang, "Detecting Reddit users with depression using a hybrid neural network SBERT-CNN," in *Proc. IEEE 11th Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2023, pp. 193–199.

[136] H. A. Sweety, D. A. Mahmud, A. Hossain, and N. A. A. Rahman, "An efficient approach to analysis sentiment on social media data using bi-long short time memory network," in *Proc. Int. Joint Conf. Adv. Comput. Intell.* Springer, Jan. 2024, pp. 583–592.

[137] G. Abercrombie and R. Batista-Navarro, "ParlVote: A corpus for sentiment analysis of political debates," in *Proc. 12th Lang. Resour. Eval. Conf.*, Mar. 2020, pp. 5073–5078.

[138] N. A. Semary, W. Ahmed, K. Amin, P. Pławiak, and M. Hammad, "Improving sentiment classification using a RoBERTa-based hybrid model," *Frontiers Hum. Neurosci.*, vol. 17, Dec. 2023, Art. no. 1292010.

[139] S. Kusal, S. Patil, A. Gupta, H. Saple, D. Jaiswal, V. Deshpande, and K. Kotecha, "Sentiment analysis of product reviews using deep learning and transformer models: A comparative study," in *Proc. Int. Conf. Artif. Intell. Textile Apparel*. Springer, Jan. 2024, pp. 183–204.

[140] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019.

[141] Z. Ye, T. Zuo, W. Chen, Y. Li, and Z. Lu, "Textual emotion recognition method based on ALBERT-BiLSTM model and SVM-NB classification," *Soft Comput.*, vol. 27, no. 8, pp. 5063–5075, Apr. 2023.

[142] T. Elghazaly, A. Mahmoud, and H. A. Hefny, "Political sentiment analysis using Twitter data," in *Proc. Int. Conf. Internet Things Cloud Comput.*, Mar. 2016, pp. 1–5.

[143] A. Nanayakkara and G. Thennakoon, "Sentiment analysis of YouTube comments using deep neural networks and pre-trained word embedding," *IUP J. Comput. Sci.*, vol. 17, no. 3, 2023.

[144] H. D. Huynh, H. T.-T. Do, K. Van Nguyen, and N. L.-T. Nguyen, "A simple and efficient ensemble classifier combining multiple neural network models on social media datasets in Vietnamese," 2020, *arXiv:2009.13060*.

[145] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using BERT," in *Proc. Artif. Intell. Transforming Bus. Soc. (AITB)*, vol. 1, Nov. 2019, pp. 1–5.

[146] S. Brownfield and J. Zhou, "Sentiment analysis of Amazon product reviews," in *Software Engineering Perspectives in Intelligent Systems*. Berlin, Germany: Springer, 2020, pp. 739–750.

[147] A. Singh and G. Jain, "Sentiment analysis of news headlines using simple transformers," in *Proc. Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2021, pp. 1–6.

[148] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM model for document-level sentiment analysis," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 3, pp. 832–847, Jul. 2019.

[149] A. Thalange, S. Kondekar, S. Phatate, and S. Lande, "Social media sentiment analysis using the lstm model," in *Evolutionary Computing and Mobile Sustainable Networks*. Berlin, Germany: Springer, 2022, pp. 123–137.

[150] M. K. Shaik Vadla, M. A. Suresh, and V. K. Viswanathan, "Enhancing product design through AI-driven sentiment analysis of Amazon reviews using BERT," *Algorithms*, vol. 17, no. 2, p. 59, Jan. 2024.

[151] A. Kuila and S. Sarkar, "Deciphering political entity sentiment in news with large language models: Zero-shot and few-shot strategies," 2024, *arXiv:2404.04361*.

[152] T. T. Aurpa, R. Sadik, and M. S. Ahmed, "Abusive Bangla comments detection on Facebook using transformer-based deep learning models," *Social Netw. Anal. Mining*, vol. 12, no. 1, p. 24, Dec. 2022.

[153] C. Suhaeni and H.-S. Yong, "Mitigating class imbalance in sentiment analysis through GPT-3-generated synthetic sentences," *Appl. Sci.*, vol. 13, no. 17, p. 9766, Aug. 2023.

[154] J. Bodapati, N. Veeranjaneyulu, and S. Shaik, "Sentiment analysis from movie reviews using LSTMs," *Ingénierie des systèmes d Inf.*, vol. 24, no. 1, pp. 125–129, Apr. 2019.

[155] A. Stipiuc, "Romanian media landscape in 7 journalists' Facebook posts: A ChatGPT sentiment analysis," *Saeculum*, vol. 57, no. 1, pp. 20–46, Jul. 2024.

[156] T. Srivastava, D. Arora, and P. Sharma, "Sentiment analysis of COVID-19 tweets using BiLSTM and CNN-BiLSTM," in *Proc. Int. Conf. Recent Trends Comput.* Springer, 2023, pp. 523–535.

[157] X. Zhao and C.-W. Wong, "Automated measures of sentiment via transformer- and lexicon-based sentiment analysis (TLSA)," *J. Comput. Social Sci.*, vol. 7, no. 1, pp. 145–170, 2024.

[158] U. M. Ramirez-Alcocer, E. Tello-Leal, J. D. Hernandez-Resendiz, and G. Romero, "A hybrid CNN-LSTM approach for sentiment analysis," in *Proc. Congr. Intell. Syst.* Springer, Jan. 2024, pp. 425–437.

[159] F. Hamborg, K. Donnay, and B. Gipp, "Towards target-dependent sentiment classification in news articles," in *Proc. 16th Int. Conf.*, Beijing, China. Springer, Jan. 2021, pp. 156–166.

[160] G. Yenduri, M. Ramalingam, G. C. Selvi, Y. Supriya, G. Srivastava, P. K. R. Maddikunta, G. D. Raj, R. H. Jhaveri, B. Prabadevi, W. Wang, A. V. Vasilakos, and T. R. Gadekallu, "GPT (generative pre-trained transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *IEEE Access*, vol. 12, pp. 54608–54649, 2024.

[161] A. Alhadlaq and A. Altheneyan, "DistilRoBERTa2GNN: A new hybrid deep learning approach for aspect-based sentiment analysis," *PeerJ Comput. Sci.*, vol. 10, p. e2267, Aug. 2024.

[162] A. Ahmet and T. Abdullah, "Real-time social media analytics with deep transformer language models: A big data approach," in *Proc. IEEE 14th Int. Conf. Big Data Sci. Eng. (BigDataSE)*, Dec. 2020, pp. 41–48.

[163] A. Müller, J. Riedl, and W. Drews, "Real-time stance detection and issue analysis of the 2021 German federal election campaign on Twitter," in *Proc. Int. Conf. Electron. Government*. Springer, 2022, pp. 125–146.

[164] S. Hussain, N. Dhanda, and R. Verma, "Sentiment analysis of Amazon product reviews using VADER and RoBERTa models," in *Proc. 8th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jun. 2023, pp. 708–713.

[165] S. Eyvazi-Abdoljabbar, S. Kim, M.-R. Feizi-Derakhshi, Z. Farhadi, and D. Abdulameer Mohammed, "An ensemble-based model for sentiment analysis of Persian comments on Instagram using deep learning algorithms," *IEEE Access*, vol. 12, pp. 151223–151235, 2024.

[166] J. Biswas, M. M. Rahman, A. A. Biswas, M. A. Khan, A. Rajbongshi, and H. A. Niloy, "Sentiment analysis on user reaction for online food delivery services using BERT model," in *Proc. 7th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2021, pp. 1019–1023.

[167] K. Quoc Tran, A. Trong Nguyen, P. G. Hoang, C. D. Luu, T.-H. Do, and K. Van Nguyen, "Vietnamese hate and offensive detection using PhoBERT-CNN and social media streaming data," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 573–594, Jan. 2023.
[168] K. L. Tan, C. P. Lee, K. M. Lim, and K. S. M. Anbananthen, "Sentiment analysis with ensemble hybrid deep learning model," *IEEE Access*, vol. 10, pp. 103694–103704, 2022.
[169] J. Soni and K. Mathur, "Enhancing sentiment analysis via fusion of multiple embeddings using attention encoder with LSTM," *Knowl. Inf. Syst.*, vol. 66, no. 8, pp. 4667–4683, Aug. 2024.
[170] S. Zhang, H. Yu, and G. Zhu, "An emotional classification method of Chinese short comment text based on ELECTRA," *Connection Sci.*, vol. 34, no. 1, pp. 254–273, Dec. 2022.
[171] F. Wu, Z. Shi, Z. Dong, C. Pang, and B. Zhang, "Sentiment analysis of online product reviews based on SenBERT-CNN," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, Dec. 2020, pp. 229–234.

AISH ALBLADI received the M.S. degree in computer science from Ball State University, Muncie, IN, USA. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA. His research interests include AI, natural language processing, and sentiment analysis.

MINARUL ISLAM received the B.S. degree from the Department of Computer Science and Engineering, Jessore University of Science and Technology, Jessore, Bangladesh, in 2016, and the M.S. degree from the Department of Electrical and Electronic Engineering, Universiti Malaysia Pahang, Pahang, Malaysia, in 2021. Currently, he is pursuing the full-time Ph.D. degree with the Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA. He has published more than 11 research papers at different conferences and peer-reviewed journals. Recently, his current project poster abstract has been accepted at the ACM SenSys 2024 Conference, which is one of the top conference in the area of mobile computing. His primary research interests include machine learning, mobile sensing, and wireless sensor networks. During his M.S. degree, he was awarded the Bronze, Silver, Gold, and Best Innovative Technology Awards.

CHERYL SEALS is currently an Associate Professor with the Department of Computer Science and Software Engineering, Auburn University. Her research interests include human–computer interaction, user interface design, usability evaluation, and educational gaming technologies. She also works with outreach initiatives to improve computer science education at all levels. The programs are focused on increasing the computing pipeline by getting students interested in STEM disciplines and future technology careers.