A Review of Deep Learning Models for Twitter Sentiment Analysis: Challenges and Opportunities

Laxmi Chaudhary[1], Nancy Girdhar[2,*], Deepak Sharma[3,*], Javier Andreu-Perez[4,*], Antoine Doucet[5], and Matthias Renz[6]

[1]*Dept. of Computer Science & Engineering, Jaypee Institute of Information Technology, Noida, India*
[2,5]*L3i, University of La Rochelle, La Rochelle, France*
[3,6]*Dept. of Computer Science, Christian-Albrechts-University, Kiel, Germany*
[4]*School of Computer Science & Electronic Engineering, University of Essex Colchester, United Kingdom*
[*]*Corresponding author: Nancy Girdhar, nancy.gr1991@gmail.com; Deepak Sharma, deepak.btg@gmail.com, j.andreu-perez@essex.ac.uk*

***Abstract*—Microblogging site Twitter is one of the most influential online social media websites, that offers a platform for the masses to communicate, express their opinions, and share information on a wide range of subjects and products, resulting in the creation of a large amount of unstructured data. This has attracted significant attention from researchers, who seek to understand and analyze the sentiments contained within this massive user-generated text. The task of sentiment analysis entails extracting and identifying user opinions from the text, and various lexicon and machine learning-based methods have been developed over the years to accomplish this. However, deep learning-based approaches have recently become dominant due to their superior performance. The current study briefs on standard preprocessing techniques and various word embeddings for data preparation. It then delves into a taxonomy to provide a comprehensive summary of deep learning-based approaches. Additionally, the work compiles popular benchmark datasets and highlights evaluation metrics employed for performance measures as well as the resources available in the public domain to aid sentiment analysis tasks. Furthermore, the survey discusses domain-specific practical applications of sentiment analysis tasks. Finally, the study concludes with various research challenges and outlines future outlooks for further investigation.**

***Index Terms*—twitter, sentiment analysis, opinion mining, deep learning, natural language processing, social network**

I. Introduction

Over the past few years, social media platforms such as Twitter, Instagram, Facebook, and various blogging sites have experienced exponential growth in their user base. These venues allow users to be more vocal about their opinions, emotions, and thoughts on diverse topics and items of their interests, resulting in the generation of a surplus multitude of data [1, 2, 3, 4, 5, 6]. Moreover, besides textual content, the various aspects of multi-modality include pictures, audio, and video, which has piqued the interest of the research community to identify, extract, and analyze user sentiments exhibited in the text, referred to as Sentiment Analysis (SA). Among social networking sites, Twitter, with over 330 million active microblogging service users, has become a popular source of data for sentiment analysis due to its real-time nature and the sheer volume of data generated [7]. The analysis of user-generated content is crucial for various business applications, as it provides insights into users' daily lives, and explains their behavior and activities, as well as how they are influenced by others' opinions. The task of sentiment analysis can yield valuable knowledge for further detailed analysis, including identifying trends or results of a particular topic based on sentiment [8], such as movie preferences[9], product proclivity in the market [10, 11], or political opinions [12].

Despite the growing interest in SA, classifying the sentiment polarity of Twitter tweets remains a crucial task due to several factors, including language and the lack of contextual cues. Such factors may contradict the well-formed language embodied in most corpora used for text analysis. Therefore, there is an increasing interest in improving sentiment classification methods to achieve more accurate, explainable, and traceable outcomes, as well as better performance in real-time applications. Numerous studies have been conducted to improve sentiment analysis techniques, as evidenced by the recent SemEval challenges [13] and there is still much work to be done to enhance sentiment classification methods further [14].

Various sentiment analysis techniques, including traditional ones such as lexicon-based methods [15, 16, 17, 18, 19, 20, 21, 22, 23], machine-learning algorithms [24, 25, 26, 27, 28, 29], and hybrid approaches, have been employed for analyzing Twitter data. Additionally, graph-based approaches have also been suggested to identify sentiment in Twitter datasets [30, 31]. However, these techniques have certain limitations, such as handling natural language complexities, short sequences of text, semantic relationships, feature selection, lack of validation results, and processing large amounts of data, which hinder their real-time applicability, especially with high-dimensional features.

To address these limitations, deep learning, a cluster of multi-layer neural network algorithms have emerged as a promising sub-field of machine learning for Twitter sentiment analysis [32, 33, 34]. Several deep learning-based models, including Deep *(Vanilla)* Neural Networks (DNN) Ali et al.

[32], Yasir et al. [34], Convolutional Neural Networks (CNN) [35, 36, 37, 38], Recurrent Neural Networks (RNN) [39, 40], and their variants such as Long Short-Term Memory (LSTM) [41, 42, 43, 44], Gated Recurrent Units (GRU) and hybrid techniques have shown effectiveness in capturing the nuances of natural language and handling the noise and ambiguity present in Twitter data [35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. These models offer flexible solutions that enhance sentiment analysis performance by providing a better interpretation of the context and semantic meaning of text data.

Motivation-***Twitter as a Unique Case for Sentiment Analysis***: Twitter presents a distinctive environment for sentiment analysis, characterized by the specific features that set it apart from other contexts. Firstly, the stringent character limit, which ranges from 280 to 10,000 characters (depending on the subscription) per tweet, leads to concise expressions. This can result in the loss of nuanced sentiment cues. Additionally, Twitter users often employ informal language, slang, and abbreviations, posing challenges for sentiment analysis algorithms to comprehend unconventional language usage accurately. Moreover, the widespread use of emojis and hashtags in tweets requires specialized techniques to effectively integrate these non-textual elements into sentiment analysis, capturing their emotive context.

Furthermore, Twitter data is often noisy, with promotional content, news updates, and irrelevant information intermingled with sentiment expressions. This noise hampers sentiment analysis performance and necessitates robust preprocessing techniques to filter out irrelevant content and enhance sentiment prediction accuracy. The challenges posed by Twitter sentiment analysis, such as handling brevity, informality, non-textual cues, and noise, demand tailored preprocessing strategies and algorithms to ensure reliable sentiment analysis results.

Given the significance of sentiment analysis in a vast spectrum of applications, and the plethora of work dedicated to sentiment analysis within Twitter literature, various lines of review studies are presented by the researchers in order to highlight the advances being achieved and the challenges yet need to be addressed. Mittal and Patidar [45] focused on exploring lexicon-based and machine learning-based methods for sentiment analysis on Twitter. However, this study did not delve extensively into preprocessing methods and deep learning techniques. In contrast, Silva et al. [46] conducted a comprehensive survey primarily centered around semi-supervised approaches, encompassing graph-based, wrapper-based, and topic-based methods for tweet classification. This survey featured a comparative analysis of three semi-supervised techniques: self-training, co-training, and topic modeling. Azzouza et al. [47] introduced a system aimed at discovering and tracking opinions on Twitter using Apache Storm. Through dynamic graphical visualizations, multiple opinions were represented, while an unsupervised machine-learning technique was employed for sentiment analysis and polarity detection, and the evaluation of the model's performance was conducted using SemEval datasets. Additionally, Ligthart et al. [48], Wankhade et al. [49], and Das and Singh [50] have conducted studies to gain insights into diverse tasks and approaches within sentiment analysis.

Some studies have summarized the technical and theoretical aspects of sentiment analysis, as done by Yadav and Vishwakarma [51] and Sharma and Jain [52]. Others have compiled literature to address the challenges posed by large data and the expansion of sentiment analysis into domains like marketing, finance, healthcare, and disaster analysis. Works by De Albornoz et al. [10], Soni and Sharaff [53], and Fadel and Cemil [27] contribute to this effort. Furthermore, few researchers have investigated the impact of data quality on sentiment analysis performance, considering factors such as readability, subjectivity, and informativeness. Kumar et al. [21] and Jain and Vaidya [54] examined online product reviews to analyze customer feedback for applications like business monitoring and brand management.

In a different line of work, comparative studies, most reviews have focused on reliability metrics such as F1-score or overall accuracy, and performance evaluation of methods is often carried out on small datasets [55]. These studies have shed light on domain-specific past literature or compared the performances of different models on sentiment analysis tasks. However, despite the recent surge in deep learning-based developments in Twitter sentiment analysis, there is still a gap in the literature for an extensive analysis and outline of research progress over the years. To bridge this gap, this comprehensive study presents an objective overview of various sentiment analysis methods, with a focus on deep learning approaches, to provide an overview of existing research and identify research gaps, paving the way for researchers to fill those gaps.

Observing prevalent previous, current, and coming trends & developments, this research aims to achieve the following objectives:

- **Systematic Taxonomy:** To present a systematic taxonomy that summarizes, compares, and reviews representative works for each type of approach. This provides new perspectives for future exploration and practices in sentiment analysis.
- **Pre-processing Techniques:** To provide a summary of various pre-processing techniques used to clean and process text data before applying deep learning models. To discuss the impact of these techniques on the accuracy of sentiment analysis and their effectiveness in handling noisy and ambiguous data.
- **Overview of Traditional SA Techniques:** To provide an overview of traditional sentiment analysis techniques and their limitations in processing large volumes of Twitter data. To discuss the challenges faced by traditional techniques and the need for more advanced techniques to handle the complexities of natural language.
- **Analysis of Deep Learning-Based Approaches:** To provide a detailed analysis of various deep learning-based approaches for sentiment analysis on Twitter, including DNN, CNN, RNN, and their variants, such as LSTM and GRU. To discuss their architectures, training methodologies, and their strengths and limitations.
- **Challenges Faced in SA on Twitter:** To analyze the challenges faced in sentiment analysis on Twitter, such

as noisy data, sarcasm, and irony. To discuss the impact of these challenges on the accuracy of sentiment analysis and the need for pre-processing techniques to clean text data.
- **Performance Measures:** To present a detailed analysis of the evaluation metrics used to evaluate the performance of the models.
- **Real-world Case Studies:** To discuss various dimensions of sentiment analysis usage, its applicability, and its influence on various business domains.
- **Future Perspectives:** Finally, to provide an overview of the future research directions in Twitter sentiment analysis and the need for more robust models that can handle the complexities of natural languages and the challenges faced in processing large volumes of data.

The remainder of this article is structured as follows: Section II provides the fundamental concepts of pre-processing and word embedding. Next, Section III details the development of Twitter sentiment analysis and a review of existing literature. Section IV summarizes various available data sources, evaluation metrics, and tools. Then, Section V presents domain-specific case studies and applications of Twitter sentiment analysis. Section VI highlights various research gaps and future perspectives and finally, Section VII concludes the present study.

II. Sentiment analysis on Twitter

This section provides information about the standard preprocessing steps, and different word embeddings used to perform sentiment analysis tasks on the Twitter dataset.

A. Pre-processing of Twitter data

The input data quality significantly impacts the performance of the sentiment analysis models. The datasets that are used for sentiment analysis are often unstructured or semi-structured, containing a huge amount of irrelevant data that is not useful for predicting sentiments. For instance, when dealing with large datasets, the computational training time can be lengthy and the presence of stop-words can negatively impact the accuracy of the model. Therefore, it is necessary to preprocess the data in order to save time during training and to increase efficiency [56]. As a consequence, preprocessing text plays a crucial role in noise reduction, and data quality improvement, which further elevates the model performance. Based on our literature review, we have compiled the standard preprocessing steps adopted for Twitter SA in the state of the art as illustrated in Table I.

- **Data Collection:** To collect the relevant tweets using the Twitter API[1] or other tools[2][3][4].
- **Data Cleaning:** To remove any irrelevant information from the tweets, such as URLs, usernames, hashtags, special characters, and numbers.
- **Data Balancing:** To have a balanced dataset with equal representation of each sentiment class. Therefore, techniques such as oversampling, under-sampling, or data augmentation are used to balance the data.
- **Tokenization:** To split the cleaned tweets into individual words or tokens. This involves breaking down a text into tokens such as words, numbers, punctuation marks, etc [56]. This is done to prepare the data for further processing.
- **Stop-Word Removal:** Stop-words are frequently occurring words, such as “an”, “in”, “of”, “a”, “is”, “the”, “to” etc. However, they do not add much value to text analysis. Hence, are removed to reduce the noise in the data and to improve the efficiency of sentiment analysis [56].
- **Stemming:** Stemming is also termed as the text standardization where the tokens are truncated to their root form to reduce feature complexity and enhance the learning capability of classifiers [56].
- **Lemmatization:** It is a similar process to stemming but with a predefined dictionary that retains the context of the word and ensures that the meaning is not lost [56].
- **Short-Word Removal:** Remove words less than three characters to enhance the accuracy and robustness of classifiers [56].
- **Case Conversion:** Convert text into lowercase to avoid any case-sensitivity issues that could affect the classifier’s performance [56].
- **Punctuation Removal:** Remove punctuation marks from the text such as full stops, commas, brackets, etc. [57].
- **URLs Removal:** URLs are references to web locations that do not provide any additional details and are removed using regular expression matching operations [57].
- **Expanding Contractions:** Contractions like “cannot” and “do not” are often used to fit within Twitter’s character limit of a tweet/post, and are changed to actual words to improve the accuracy of sentiment analysis [7].

TABLE I: DATA PREPROCESSING STEPS

Pre-processing Steps	Publications
Tokenization	[58, 59, 60, 32, 34, 44, 57, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73]
Stop-Word Removal	[7, 58, 59, 60, 44, 57, 61, 63, 65, 69, 70, 71, 72, 73]
Stemming	[58, 59, 60, 57, 67, 70, 71, 72, 73]
Lemmatization	[59, 63, 65, 69, 71, 72, 73]
Short-Word Removal	[7]
Case Conversion	[7, 59, 60, 32, 57, 61, 62, 65, 69, 70, 72, 73]
Punctuation Removal	[7, 58, 59, 60, 34, 44, 57, 61, 69, 71, 72, 73]
URLs Removal	[7, 59, 34, 57, 62, 64, 65, 67, 69, 71, 72, 73]
Expanding Contractions	[7, 65, 69]

B. Word Embeddings

Unlike images where the input vectors are directly generated based on pixel data (which are already numeric), it is more challenging to extract input vectors from the textual data (which are strings/characters) for neural network models. To

[1]https://developer.twitter.com/en/docs/twitter-api
[2]https://docs.tweepy.org/en/stable/index.html
[3]https://twarc-project.readthedocs.io/en/latest/
[4]https://github.com/thepanacealab/SMMT

deal with this, word embedding, which is the conversion of the vocabulary of words into a vector representation is used. One-hot encoding, which is a common representation, assigns a $\mid V \mid$ dimensional vector space to each word, where $\mid V \mid$ represents the size of the vocabulary. The vector space consists of only one non-zero entry that corresponds to the word, while the rest of the entries are zeros. However, this method has some drawbacks, such as high computational requirements and the inability to handle context similarity because each word is encoded as a sparse, high-dimensional vector. The other approach is the *Term Freqency-Inverse Document Frequency* (TF-IDF) method which assigns a score that reflects the relevance of the term in the document compared to the rest of the corpus. Though it is simple, effective, and computationally efficient to identify important words in a document, and can be used to rank documents based on their relevance to a query. Nevertheless, it does not take into account the order of words in a document or their semantic meaning, and may not perform well on documents with highly specialized vocabulary or uncommon words [26, 25, 22]. An alternative approach is to use *dense embedding vectors* to obtain the context of words in terms of both syntax and semantics [74, 75, 76]. The dense vector mapping ensures that words with similar meanings are represented close to each other in the vector space. To improve the generalization, representation, and computational time of sentiment classification models, various word embeddings have been proposed. Word embeddings are a popular technique for representing textual data into numeric input vectors, that are easily processed by neural network models. Several types of word embeddings exist in the literature and some of the commonly used word embeddings are summarized as follows:

- **Word2Vec:** *Word2Vec* [77] is a popular word embedding model that is based on neural networks which are designed to reconstruct the linguistic contexts of the words [74, 78, 79]. It employs a two-layer neural network architecture that takes text as input and generates a vector embedding for each word as output. There are two types of *Word2Vec* models: *Skip-gram* and *Continuous Bag of Words* (CBOW). The *Skip-gram* model predicts a D-dimensional vector representation of each word in the corpus. The input and hidden layers have the same number of neurons as the vocabulary size and the word vector dimensions, respectively. The weights between these layers are represented by $W_{H\times D}$, where H is the size of the hidden layer. This weight matrix signifies the likelihood of each word's occurrence for that input. The model learns the correlation between words in a vocabulary by computing the error at the output layer using a loss function and updating the weights (word embeddings) through backpropagation. In contrast, the CBOW model processes the context of a word as input and predicts the word based on that context [74, 79].
- **GloVe** or *Global Vectors* **:** *Glove* [80] is another popular word embedding model that is based on co-occurrence statistics. It is an unsupervised learning method that learns word embeddings by factorizing a matrix of word co-occurrence probabilities and combines the local context window and matrix factorization methods to analyze the local and global statistics of a corpus [80]. It performs better in apprehending the analogy of words than matrix factorization (a.k.a *Latent Semantic Analysis*) which only generates an efficient substructure of vector space. On the other hand, the Skip-gram (local context window approach) performs well on analogy tasks but does not make full use of the corpus statistics [74, 81].
- **FastText:** *FastText* [82] is an extension of Word2Vec that represents each word using n-grams of characters instead of individual words [83]. It learns word embeddings by representing each word as a bag of characters and then learning embeddings for these n-grams. This allows the model to generate efficient embeddings of rarely occurring words in the corpus. The n-grams are employed to train a Skip-gram model, and the embedding of a word is determined by summing up the embeddings of all its n-grams. However, it requires high memory and system requirements to create embeddings of each character n-gram in the vocabulary [74, 84]. FastText has been shown to improve the performance of sentiment analysis models on Twitter data, especially for *out-of-vocabulary* words.

Besides the aforementioned three popular word embedding schemes (*Word2Vec*, *GloVe*, and *FastText*), other word-embedding approaches are also developed such as *BERT* which is a pre-trained language model that has been fine-tuned for sentiment analysis on Twitter data and has achieved superior performance compared to other models. *ELMo* [85] is another pre-trained language model that uses a bi-directional LSTM architecture to learn contextualized word embeddings and has shown state-of-the-art results, especially for sentiment classification at the sentence level.

III. DEVELOPMENT OF SENTIMENT ANALYSIS AND LITERATURE CLASSIFICATION

This section elaborates on the recent advances in the field of Twitter sentiment analysis. For this study, we have considered publications of the Scopus database from 2010∼2022. The Scopus document search string in the current study was composed as follows - (TITLE-ABS-KEY("twitter" AND "sentiment" AND "deep" AND "learning") AND (LIMIT-TO (LANGUAGE, "English"))) which resulted in 1115 research papers. The section highlights the year-wise publications, top organizations, key researchers, and prominent source titles of this domain. Furthermore, a detailed survey is presented on Twitter sentiment analysis, bifurcated into conventional and deep learning-based approaches along with hybrid techniques.

A. Recent Trends and Developments

- ***Annual Trends:*** Figure 1 displays the annual trend of research publications on sentiment analysis using Twitter data from 2010∼2022. The x-axis represents the publication years, while the y-axis indicates the publication count recorded in the Scopus database. The data shows a gradual increase in the number of publications from 2010∼2016, followed by a dip in 2017. However, 2018 and 2019 witnessed a significant surge in publication

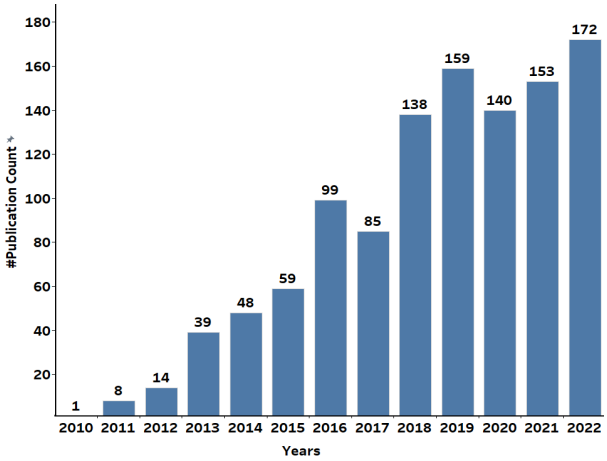

Fig. 1: Annual Publication Count.

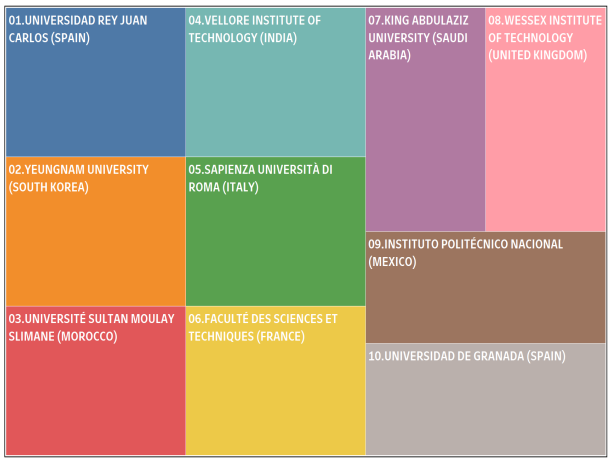

Fig. 2: Top Organizations.

counts compared to previous years, with a remarkable 19-fold increase in 2019 as compared to 2011. Despite a decline in 2020, the trend has shown an upward trend from 2020∼2022, suggesting a sustained interest in Twitter sentiment analysis research over time. Specific publication counts for each year are mentioned above the corresponding bars in the chart.

- ***Key Organizations:*** Figure 2 presents an analysis of the top ten organizations that published the most Twitter sentiment analysis-related articles in the Scopus database from 2010∼2022. Universidad Rey Juan Carlos in Spain published the highest number of 16 articles on Twitter sentiment analysis during this period, followed by Yeungnam University in South Korea with 13 articles. Université Sultan Moulay Slimane in Morocco published a total of 12 articles, while Vellore Institute of Technology in India and Sapienza Università di Roma in Italy had 11 publications each. The researchers at Faculté des Sciences et Techniques in France contributed 9 articles. King Abdulaziz University in Saudi Arabia, Wessex Institute of Technology in the UK, Instituto Politécnico Nacional in Mexico, and Universidad de Granada in Spain published 8 articles each. This analysis provides a detailed overview of the affiliations of authors worldwide who have made significant contributions to Twitter sentiment analysis research in the past twelve years.
- ***Key Authors:*** Table II presents the top ten researchers globally who have published the most articles on Twitter sentiment analysis in Scopus, along with their respective organizations. The leading author in terms of publication count is Mohammed Erritali from Sultan Moulay Slimane University in Beni Mellal, Morocco, with a total of 9 articles. The subsequent five top authors come from organizations in Italy, the UK, and Spain, and have contributed 8 articles each. Furthermore, Ana Reyes-Menendez, affiliated with Rey Juan Carlos University in Spain, has published 7 articles. The remaining three authors, affiliated with organizations in South Korea, Morocco, and Spain, have published 6 articles each. This analysis highlights the presence of highly productive research groups focusing on Twitter sentiment analysis in Spain.
- ***Key Sources:*** Figure 3 presents an analysis of the top ten source titles that published articles on Twitter sentiment analysis between 2010∼2022. The Lecture Notes in Computer Science, which includes the subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, had the highest 72 publications on Twitter sentiment analysis. The second-highest number of publications, 37, came from Advances in Intelligent Systems and Computing, followed by Communications in Computer and Information Science with 35 articles, and the ACM International Conference Proceeding Series with 31 articles. Ceur Workshop Proceedings and Technology and IEEE Access had 28 and 22 publications, respectively. Additionally, Social Network Analysis and Mining and the International Journal of Advanced Computer Science and Applications published 15 and 12 articles each. Finally, Lecture Notes in Networks and Systems and Procedia Computer Science were ranked ninth and tenth, with 11 publications each, respectively. This analysis provides detailed information on the significant source titles that have contributed to the research in this domain.

- ***Global View:*** Figure 4 displays the relative percentage of publications according to the article count per country. It is evident from the figure that a majority of the publications on Twitter sentiment analysis have first authors from Asia, followed by the USA and Europe. India has the highest number of publications in this field, with 222 articles, indicating a considerable research interest in Twitter sentiment analysis among Indian authors. United States researchers have the second-highest number of publications, with 145 articles. Other countries such as Spain, Italy, China, the UK, Brazil, Saudi Arabia, South Korea, and Morocco have also contributed a significant amount to this field, providing a decent level of diversity. Furthermore, researchers from Egypt, Mexico, Pakistan, Australia, Germany, Turkey, Japan, Iran, Malaysia, and Canada have also made a noteworthy contribution to this

TABLE II: KEY AUTHORS (**P**: Publication, **C**: Citation, **AC**: Average Citation)

S.No.	Author	P	C	AC	Organization
1	Mohammed Erritali	9	123	13.67	Sultan Moulay Slimane University, Beni Mellal, Morocco
2	Francesco Borghini	8	21	2.63	Safety Security Engineering Group–DICMA, Sapienza University of Rome, Italy
3	Fabio Grazia	8	21	2.63	Wessex Institute of Technology, Southampton, United Kingdom
4	Mara Lombardi	8	22	2.75	Safety Security Engineering Group–DICMA, Sapienza University of Rome, Italy
5	Soodamani Ramalingam	8	21	2.63	School of Physics, Engineering and Computer Sciences, University of Hertfordshire, Hatfield, United Kingdom
6	Jose Ramon Saura	8	112	14	Department of Business Economics, Rey Juan Carlos University, Madrid, Spain
7	Ana Reyes-Menendez	7	109	15.57	Department of Business Economics, Rey Juan Carlos University, Madrid, Spain
8	Dosam Hwang	6	49	8.17	Department of Computer Engineering, Yeungnam University, Gyeongsan, South Korea
9	Youness Madani	6	52	8.67	Faculty of Sciences and Technics, Sultan Moulay Slimane University, Beni Mellal, Morocco
10	Rafael Valencia-García	6	43	7.17	Department of Computing and Systems, University of Murcia, Murcia, Spain

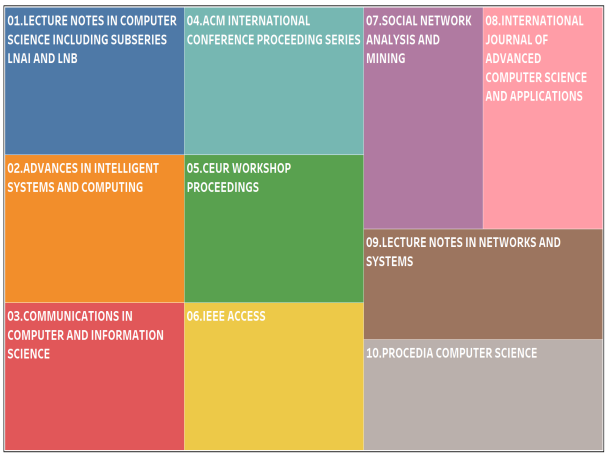

Fig. 3: Top Sources.

Fig. 4: Top Countries.

area of research.

Figure 5 showcases the visual representation of the trending keywords for sentiment analysis. The word cloud is based on the author's mentioned keywords in scientific publications that spotlight key themes and topics of interest of various researchers in this field.

B. Literature Survey

Sentiment analysis has drawn significant attention from the research community and emerged as a topic of interest, and thus, surfeit approaches and techniques are proposed to address this task. In this section, we partition diverse approaches

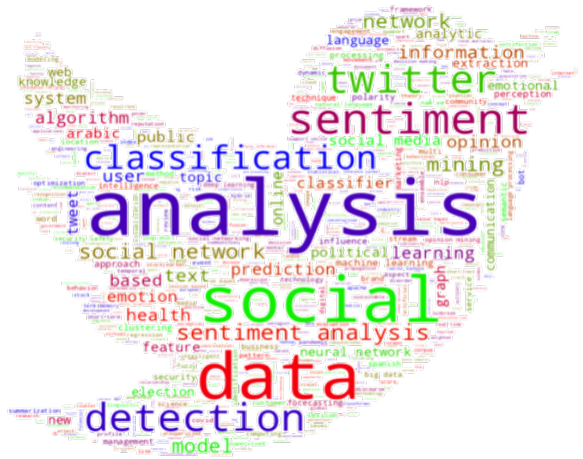

Fig. 5: Keywords Word Cloud.

proposed in the literature into two broad categories: *Conventional* and *Deep-Learning*. The first category, *conventional approaches* include methods based on *lexicons* and *machine learning*, and the latter is based on *deep neural network* models. Figure 6 depicts the evolution of sentiment analysis techniques, from lexicon-based methods to deep learning methods, and Figure 7 categorizes the publication counts of trends from various approaches during the given survey period.

1) Lexicon-based approaches: Popularly known as rule/corpus-based approaches, rely on pre-defined dictionaries or word lists with assigned polarity scores to determine the sentiment of a given dataset without any training. One of the earliest and most widely used sentiment lexicons is the *SentiWordNet* [86]. Other popular lexicons include the *AFINN* [87], the *VADER* [88], *wordnet*, and *q-word*, which are used by researchers to match words from the input statement [89]. Many studies have utilized lexicon-based approaches, such as Jurek et al. [90] which developed a sentiment analysis algorithm that focuses on real-time analysis of Twitter content. This method includes two main components: a combination function based on evidence and sentiment normalization, which are used to estimate the sentiment intensity. Table III provides further details on lexicon-based state-of-the-art.

Lexicon-based approaches can be divided into two subcategories: *Dictionary-based* and *Corpus-based*. *Dictionary-based* approaches use predefined dictionaries for instance *SentiWordNet*, and *WordNet* to perform sentiment analysis [18, 19, 20, 21]. *Corpus-based* approaches leverage corpus

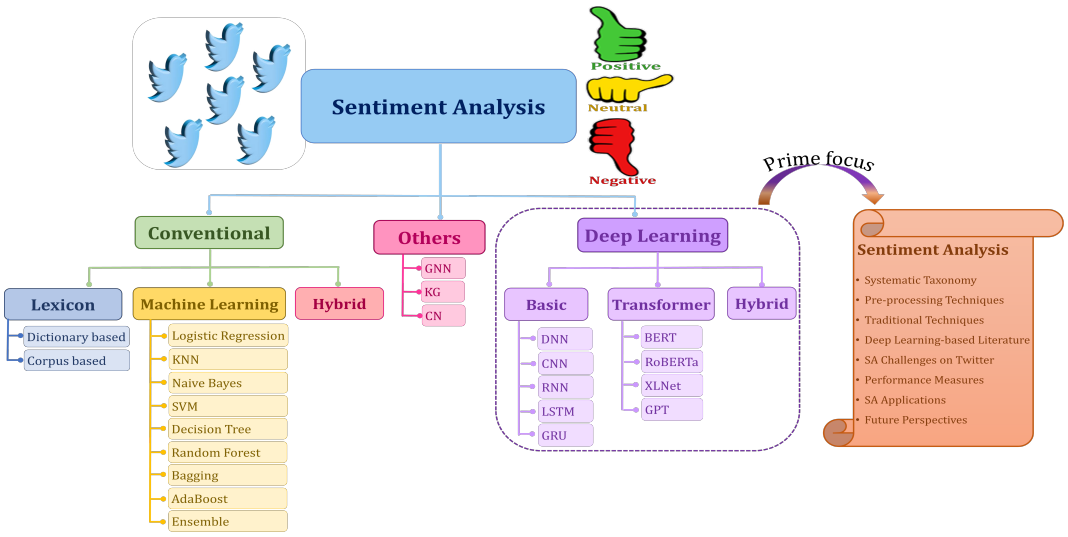

Fig. 6: Taxonomy and Objectives of Deep Learning-based Twitter Sentiment Analysis.

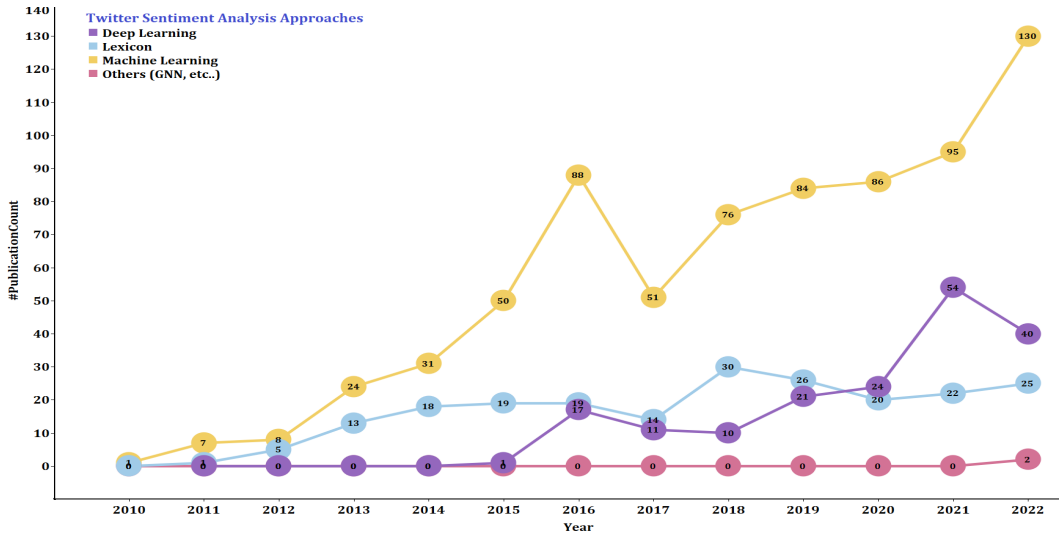

Fig. 7: Publication Counts of Trends Across Various Approaches (2010∼2022).

data for sentiment classification that are further divided into *statistical* and *semantic* sub-categories [24]. The statistical category comprises Conditional Random Field (CRF) [91], K-Nearest Neighbors (KNN) [92], and Hidden Markov Models (HMM) [53] etc.

Al-Khalisy and Jehlol [19] proposed a dictionary-based approach for extracting significant information from terrorist propaganda such as account name, location, and supporter data. This method utilized bag-of-words (BOW) to compute the overall scores for each tweet that represents the training data and to analyze the polarity, the created word list comprised of antonyms and synonyms from the dictionary. Chalothorn and Ellman [15] suggested the use of lexical resources such as NLTK toolkit, SentiWordNet, and WordNet for the analysis of online radical posts. The polarity and text intensity are calculated to analyze the sentiment. For this, the text corpus was initially acquired from various web platforms like Qawem and Montada, and after essential data pre-processing, various attribute-driven measures were employed to identify and manage extremist and religious content. Based on [15], Gitari et al. [16] build their hate verb lexicon, starting with a basic verb list, and expanding it iteratively by adding synonyms and hypernyms of the seed verbs depending on WordNet relations. Simon et al. [23] developed a corpus-based approach that uses divergent behavior to analyze the sentiment of tweets during the Kenya Westgate Mall attack to find the radicalization time among the users of Twitter. The authors recommended emergency organizations and communication centers minimize the use of negative sentiments when they communicate with the public. Another corpus-based method was proposed by Mansour [22] to analyze public sentiment polarity from Eastern and Western countries towards ISIS. This method employs text sentiment analysis using TF-IDF for analyzing the frequency of words and word sentiment. Other lexicon-based method proposed by Kharde et al. [18] uses part-of-speech (POS) tagging, while the lexicon approaches

presented by Ferrara et al. [20], and Kumar et al. [21] depend on a dictionary for feature extraction from the dataset.

Numerous techniques have utilized lexical approaches as they do not require annotated data which is one of the key challenges in the sentiment analysis task. However, these methods have certain limitations such as their accuracy being influenced by the size and quality of the lexicon. Moreover, these approaches cannot handle sarcasm and irony, which are common on Twitter. Furthermore, these methods cannot handle out-of-vocabulary words, which can lead to incorrect sentiment classification. This is particularly problematic for Twitter data, which is constantly evolving, requiring frequent updates to the lexicon. Additionally, they rely on handcrafted features, which can be a laborious and time-consuming process. Another limitation of these methods is that they are not very effective at generalizing to different domains or context-specific orientations [23].

2) Machine Learning-based approaches: These techniques have been extensively employed for sentiment analysis on Twitter in recent years. Leveraging statistical techniques, these methods have the ability to automatically learn patterns and relationships from data, which are then used to classify the sentiment of the text.

Machine Learning (ML) approaches broadly comes under the umbrella of conventional methods that constitutes popular techniques such as support vector machines (SVM) [24, 98, 99], Naïve Bayes (NB) classifier [24, 25, 99] and maximum entropy classifier [55, 100] etc. These techniques have been used in several studies for sentiment analysis, including studies on terrorism [25, 101, 102], hate speech detection [103, 104], customer satisfaction [105], and sentiment polarity detection [27] For instance, Wei et al. [24] have used the Naïve Bayes algorithm to classify tweets as positive, negative, or neutral based on the presence of specific words in a tweet. Additional traditional ML approaches have also been utilized in various other studies [26, 27, 28, 29, 103, 101, 106], which are presented in detail in Table IV.

Wei et al. [24] proposed a KNN classifier-based approach for sentiment classification to identify extremist-related conversations on Twitter public tweets. Similarly, Azizan and Aziz [25] utilized the Naïve Bayes algorithm to detect extremist affiliations in social media communication. Their model classifies user reviews into positive and negative sentiments to reflect affiliations with extremist or non-extremist groups. However, this method does not consider the overall dependencies concerning a sentence in a given document. Rani and Singh [98] proposed an SVM model with features extracted using the TF-IDF method for sentiment analysis in which they detected sentiment polarity using two SVM methods and concluded that the linear SVM model outperformed the kernel SVM.

Omer [101] proposed a machine learning-based approach that collects and uses three different datasets, including supporters of ISIS, anti-supporters of ISIS, and random tweet datasets that are unrelated to ISIS. The method employs three primary classifiers, namely Naive Bayes, AdaBoost, and Support Vector Machine. Nouh et al. [103] developed a novel ML-based approach to analyze radical content and extremism propaganda in tweets. Kaati et al. [102] introduced a method for identifying the Twitter accounts of jihadist group supporters, and online propaganda propagators using feature engineering, which involves analyzing data dependencies and classifying features as data-independent or data-dependent. Ferrara et al. [20] developed a sentiment analysis technique that uses metadata as a feature, together with a greedy selection method, and applies the Random Forest classifier and Logistic Regression models to predict the extremists' sentiment polarity in interactions.

Omar et al. [104] identified the relationship between hate speech and topics present on online social platforms based on an ML method. This approach utilizes multi-label classification by employing Logistic Regression, Linear SVC, and Random Forest classifiers. To classify text sentiment into positive, neutral, or negative, the authors have utilized feature representations that include TF-IDF, N-gram, and BOW. Rehman et al. [107] have proposed a method to detect radical text on Twitter, where religious language plays a significant role in radicalization. The authors have utilized both radical and religious features for training the model and applied TF-IDF for feature engineering to feed into ML classifiers including Random Forest, SVM, and Naïve Bayes to detect the sentiment polarity.

In order to improve the accuracy of sentiment analysis, researchers have dedicated efforts to developing ML-hybrid models [26, 108, 29, 109] that integrate multiple ML approaches to address the shortcomings of individual methods. While these hybrid models have led to better results, there is still potential for further enhancement of their outcomes.

Despite the success of machine learning-based Twitter sentiment analysis approaches, there still exist challenges that need to be addressed. One of the major limitations is their dependence on the quality and size of the training dataset. If the training dataset is biased or too small, it may lead to poor performance of the model. Another limitation is their inability to handle the ambiguity and complexity of the multi-lingual dataset and their inadequacy to efficiently capture relevant features from short sequences of text (short-text). For instance, sarcasm and irony in tweets can often be misinterpreted by these models, leading to incorrect sentiment classification. Also, their performance relies on the amount of annotated data available for training, making them highly data-dependent. However, annotating tweets is a costly process due to the dynamic nature of Twitter content. Moreover, these methods are domain-specific, which means that their effectiveness is limited to the domain in which they are trained. If they are used in a different domain, their efficacy decreases and they need to be retrained in that domain to perform well [51]. Furthermore, the step of feature extraction and engineering in ML-based methods is computationally expensive and time-consuming.

Overall, the use of ML approaches for Twitter sentiment analysis has shown promising results, nevertheless, there is still room for improvement. Further address is needed to develop more robust and efficient models for sentiment analysis, which can handle the complexities of natural languages.

3) Deep Learning-based approaches: Unlike hand-crafted feature engineering in ML approaches, models based on the DL paradigm are capable of automatically extracting signif-

TABLE III: LEXICON BASED APPROACHES

[Ref] Publication [Year]	Feature Extraction	Metrics	Limitations
[93] Zhang et al. [2011]	Unigrams, POS tagging	Precision, Recall, F1-score, Accuracy	Empirical study, hence lacks generalizability.
[17] Taboada et al. [2011]	Dictionary	Precision, Recall, F1-score, Accuracy	Unable to model the semantic relationship between an aspect and its context.
[15] Chalothorn and Ellman [2012]	BOW POS	Precision, Recall, F1-score, Accuracy	Experiments conducted on significantly limited dataset.
[94] Palanisamy et al. [2013]	POS tagging	Precision, Recall	The outcomes do not reflect the semantic contextuality of tweets.
[23] Simon et al. [2014]	Dictionary	Accuracy	Ambiguous result visualization and requires in-depth analysis.
[16] Gitari et al. [2015]	Subjectivity & Theme-based	Precision, Recall, F1-score	Possibility of performance improvement with other machine learning techniques leveraging specific theme-based features.
[95] Agarwal et al. [2015]	Unigrams, Bigrams, Bi-tagged, Dependency Features	Accuracy	Handcrafted way of mining dependency features.
[20] Ferrara et al. [2016]	Dictionary	Precision, Recall, F1-score, AUC	Considered two hypotheses without any performance analysis.
[22] Mansour [2018]	TF-IDF	Accuracy	Multilingual tweets were not taken into account.
[96] El Rahman et al. [2019]	Dictionary	Precision, Recall, F1-score	Experiments conducted on limited dataset.
[97] Mashuri et al. [2019]	Dictionary, POS tagging	Precision, Recall, Accuracy	Result outcomes overlook semantic features.
BOW: Bag-of-Words; **POS**: Parts-of-Speech; **TF-IDF**: Term Frequency-Inverse Document Frequency; **AUC**: Area Under the Curve			

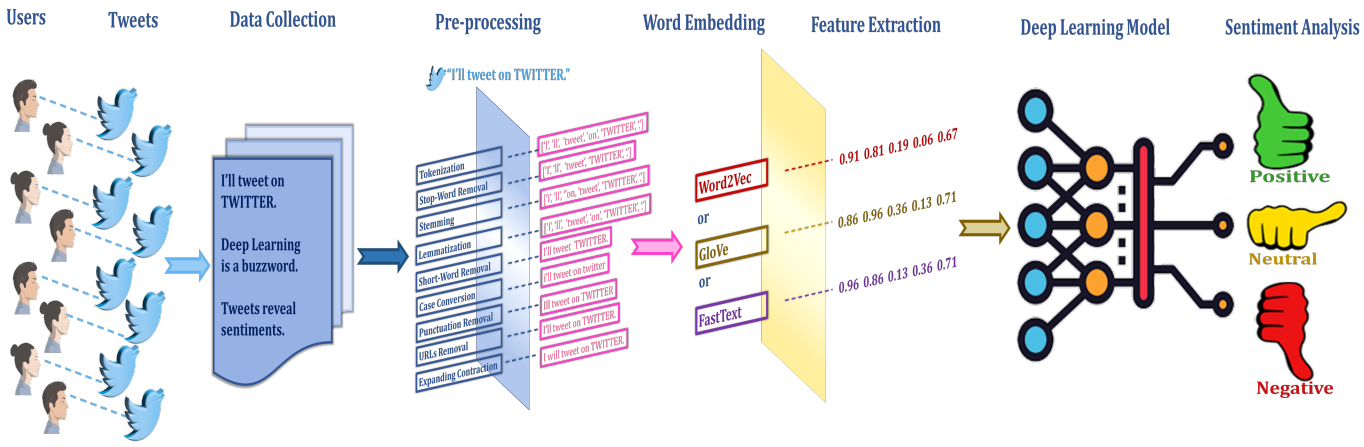

Fig. 8: Generic Pipeline of Twitter Sentiment Analysis.

icant features from the text and have shown state-of-the-art results for sentiment analysis (SA) tasks.

Deep Learning (DL) models offer several advantages over conventional methods for sentiment analysis and thus have become a recent emerging research area for Twitter sentiment analysis tasks. DL models are well-suited for handling large amounts of data that are generated every day on social media platforms. For example, on Twitter, about 6,000 tweets are produced per second on average, resulting in approximately 200 billion tweets per year. Traditional methods struggle with such surplus data, whereas DL models excel as they can learn more features while training on vast datasets, resulting in superior accuracy and performance efficiency. Additionally, deep learning models equipped with multiple hidden layers, enable them to capture complex and nonlinear patterns in the data [33] easily even in short-text data like *"tweets"*. Therefore, a plethora of DL-based models are developed over the past few decades to analyze text sentiments of posts on various social media platforms including Twitter.

To gain better insights into recent years' advancements, the current survey bifurcates the DL literature into a taxonomy broadly categorized as *Basic* and *Transformer*-based. Basic DL models consist of Deep Neural Network (DNN) [32, 33, 34], Convolutional Neural Network (CNN), [35, 36, 37, 38], Recurrent Neural Network (RNN) [40], Long Short-Term Memory (LSTM) [110] whereas Transformer-based includes BERT [111], RoBERTa [112], XLNet [113], and GPT [114] etc. Besides, these two major categories there are many DL-hybrid methods proposed by the research community for Twitter text sentiment analysis along with recent developments of *Graph-based* methods that are classified under the "other" category in the current study. The following sections detail the literature for each category of DL models. Figure 8 presents the generic DL-based pipeline for Twitter sentiment analysis.

- **Deep Neural Networks (DNNs)** are a type of artificial neural network that consists of multiple hidden layers

TABLE IV: TRADITIONAL MACHINE LEARNING AND HYBRID BASED METHODS

[Ref] Publication [Year]	Feature Extraction	Method	Metrics	Limitations
[101] Omer [2015]	Stylometry & Time-based	AdaBoost, SVM, NB	Accuracy	Performance measured on limited dataset.
[102] Kaati et al. [2015]	Contingent & non-Contingent-based	AdaBoost	Precision, Recall, Accuracy	Suboptimal results on Arabic dataset.
[20] Ferrara et al. [2016]	Greedy-based	LR, RF	Precision, Recall, F1-Score, AUC	Experiments conducted on positive polarity inclined data.
[24] Wei et al. [2016]	Sentiment Tendency, Extremism Support, Mention-Network	NB, LR, SVM, KNN	Accuracy	Biased model performance due to selective feature selection.
[26] Mirani and Sasi [2016]	TF-IDF, Geolocation-based	SVM, RF, DT, Bagging	Precision, Recall, F1-Score, Accuracy, Kappa	Experiments conducted on limited dataset.
[25] Azizan and Aziz [2017]	TF-IDF	NB	Accuracy	The overall dependencies of a sentence is not taken into account.
[106] Hartung et al. [2017]	BOW, Bi-grams	SVM	Precision, Recall, F1	Only coarse-grained features are considered.
[108] Sharif et al. [2019]	N-gram, TF-IDF	NB, SVM, DT, RF, KNN, Ensemble	Precision, Recall, F1, Accuracy	Biased detection outcomes due to weak feature selection.
[103] Nouh et al. [2019]	LIWC Dictionary, Bi-grams	SVM, KNN, NN, RF	Accuracy, Precision, Recall, F1	Utilizing bi-grams and tri-grams for binary classification resulted in low model's accuracy.
[27] Fadel and Cemil [2020]	POS Tagging	SVM, NB, LR, MV	F1, Accuracy	The negative sentiment accuracy is low and could be improved with better feature selection criteria.
[28] Smith et al. [2020]	LIWC Function	LR	Precision, Recall, F1, Accuracy, AUC	Alternative feature selection techniques may obtain more precise outcomes.
[29] Aleroud et al. [2020]	TF-IDF, LDA	SVM, KNN, DT, RF	Precision, Recall, F1	Topic modeling augmentation resulted in a disparity between the actual and anticipated model outcomes.
[104] Omar et al. [2021]	BOW, N-gram, TF-IDF	SVC, LR, RF	Precision, Recall, F1, Accuracy, Hamming Loss	Both BOW and N-grams raise false positives, impeding the model's overall efficacy.
[109] Masood and Abbasi [2021]	TF-IDF, Bi-grams	SVM, RF, LR, GNB	Accuracy, Precision, Recall, F1	The creation of the crafted dataset is not adequately elucidated.
[107] Rehman et al. [2021]	TF-IDF	NB, SVM, RF	Precision, Recall, F1, Accuracy	Reduced data samples resulted in high false positives, impacting the overall model's performance.

BOW: Bag-of-words; **POS**: Parts-of-speech; **TF-IDF**: Term Frequency-Inverse Document Frequency; **LR**: Logistic Regression; **RF**: Random Forest; **KNN**: K-Nearest Neighbor; **NN**: Neural Network; **NB**: Naive Bayes; **SVM**: Support Vector Machine; **DT**: Decision Tree; **ME**: Maximum Entropy; **AdaBoost**: Adaptive Boosting; **GNB**: Gaussian Naive Bayes; **AUC**: Area Under the Curve; **SVC**: Support Vector Classification; **LIWC**: Linguistic Inquiry and Word Count; **F1**: F1-Score

between the input and output layers (as shown in Figure 9). These textit vanilla neural networks can efficiently handle complex non-linear relationships between the layers as compared to conventional single hidden layer architectures. To perform the Twitter SA task, the DNN model implicitly learns different features from the input data in a feed-forward manner where each layer is fully connected with the next layer. During the training step, back-propagation is used to learn and adjust the weights among neurons. The weights are updated depending on the error obtained at the output layers. Ali et al. [32] developed a deep learning-based sentiment analysis model using RapidMiner to predict the results of general elections in Pakistan in 2018. Similarly, Yasir et al. [34] employed a deep learning model to forecast the interest rates of five countries, utilizing Twitter sentiments as an input. They have also integrated regression models such as linear and support vectors in their analysis.

The DNN-based models have shown superior performance compared to traditional machine learning models due to their ability to learn complex features from the data. However, these models have a larger number of hidden layers and they have a greater number of parameter values too, making them difficult to train [33].

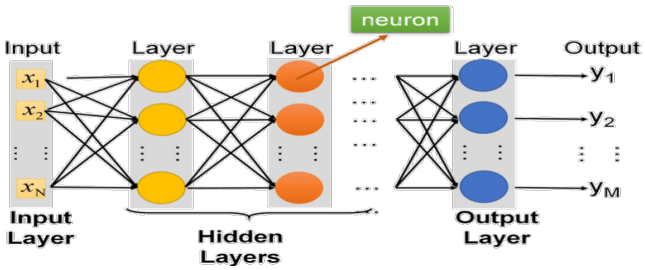

Fig. 9: DNN Architecture [115].

- **Convolutional Neural Networks (CNNs)** are one of the DNN-based model variations that typically consists of a sequence of convolutional and pooling layers, followed by one or more fully connected layers for sentiment classification. The convolutional layers use filters of

varying sizes to extract local features from the input, while the pooling layers reduce the dimensionality of the data by down-sampling the output of the convolutional layers. The extracted features are then fed into the fully connected layers for classification purposes. CNN-based models have shown to be effective for Twitter sentiment analysis and can learn complex features from the input data, allowing them to capture the context and meaning of the words. Additionally, being computationally efficient and easy to train on large datasets, these models have opted for sentiment analysis in various domains.

A generic architecture of CNN is illustrated by Figure 10. The most important component in CNN is the convolution layer. The convolutional layer h is formed by applying the activation function $f(\cdot)$ to the input matrix X, which is convolved with the weight matrix W^k and added to the bias term b_k for each layer. The elements in the i^{th} row and j^{th} column of W^k and X are referred to as $w^k_{i,j}$ and $x_{i,j}$, respectively. The resulting k feature map of h^{th} layer has a dimension $C \times H \times W$, where C, H, and W represent the channel, height, and width, respectively. One can create a convolutional layer, denoted as h, by using k small filters (also known as kernels) of size $N_i \times N_j$ as shown in Eq. 1. These filters perform a cross-correlation operation, convolving the input pixel $x_{u,v}$ to obtain $h^k_{u,v}$.

$$h^k_{u,v}(X_{u,v}) = f\left(\sum_{i=1}^{N_i}\sum_{j=1}^{N_j} w^k_{i,j} x_{u+i,v+j} + b^k\right) \quad (1)$$

Zola et al. [116] have developed a word-embedded CNN model to address cross-domain issues encountered while performing sentiment analysis. The model employs web sources such as Amazon and TripAdvisor, which contain easily labeled reviews, for fitting a sentiment prediction model. This model is later reused to classify the sentiment polarity of two unlabelled social media platforms Twitter and Facebook. The authors have also explored various techniques such as POS tagging, stemming, undersampling, oversampling, and handling unlabelled sentiment data to reduce word sparsity. Paredes-Valverde et al. [64] have proposed an approach based on Word2Vec for sentiment classification, which helps companies and organizations identify opportunities for improving the quality of their products and services.

Alharbi and de Doncker [67] have developed a CNN model that incorporates user behavioral details present in a document, such as a tweet, for sentiment analysis. The authors have utilized two datasets provided by the SemEval-2016 Workshop to evaluate the model's performance. This approach suggests that considering the content of a document or a tweet beyond its availability is advantageous in sentiment analysis, as it provides the model with an in-depth understanding of the classification task.

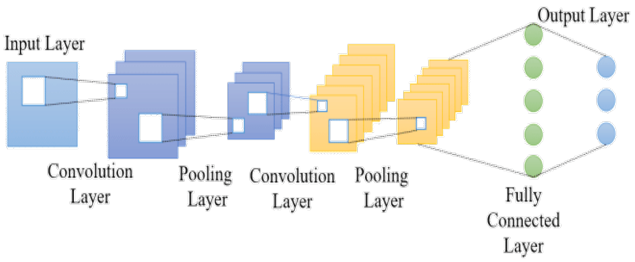

Fig. 10: CNN Architecture [117].

Overall, CNN-based models have shown to be effective for SA on Twitter data. While these models are designed to extract local features from the input and may have limitations in capturing long-term dependencies between words, recent studies have shown that incorporating attention mechanisms can improve their performance.

- **Recurrent Neural Networks (RNNs)** The main drawback of CNNs is their inability to understand the relationships between sequences. Additionally, the effectiveness of the CNN technique largely depends on choosing an appropriate window size of kernels [39]. CNN models assume that each input is unrelated to the output, which means they don't help in dealing with contextual dependencies present in the dataset. To address this, RNNs [40] aid from previous state information to handle contextual relationships to capture the temporal dependencies between words of data. These models use a hidden state that is updated at each time step, allowing them to apprehend the context and meaning of the words. Each word in the text is considered as a separate input at a given time t and previously hidden state information is employed to process the current input as presented in Figure 11.

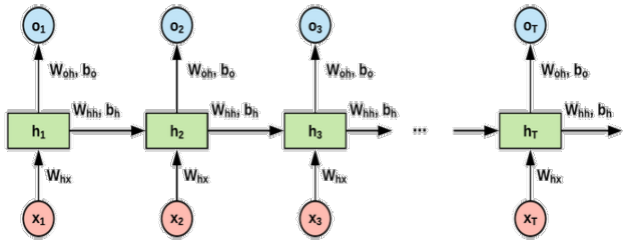

Fig. 11: RNN Architecture [118].

At a particular time step t, the input vector X_t and the output vector h_{t-1} from the preceding RNN layer are fed as inputs to the current RNN layer. The output for that time step is then computed using these two input vectors using Eq. 2.

$$h_t = tanh(h_{t-1}W_h + X_t W_X + b) \quad (2)$$

While RNNs are great at learning sequential data, they cannot obtain local attributes in parallel. As a result, RNN models are complementary to CNN models since they maintain sequential information over time. Long Short-Term Memory (LSTM) and Gated Recurrent Unit

(GRU) are extensions of RNNs widely used for sentiment analysis on Twitter.

- **Long Short Term Memory (LSTM)** As RNNs may suffer from *exploding gradient* and *vanishing gradient* issues, which makes it difficult to handle long-term contextual dependencies and fine-tune their parameters. This can lead to difficulty in training and remembering long-distance correlations in a sequential manner [39]. To resolve these issues, an LSTM a variant of RNN restructures the RNN by introducing a memory cell and a gate to retain information for further utilization and updates [110]. By modifying the RNN layer, the LSTM model solves both exploding and vanishing gradient problems occurring in RNN models. LSTM models are beneficial for sentiment classification since they can apprehend both long and short-term dependencies, and have obtained notable results in this task. Also, these models are capable to solve time-series and sequential problems with remarkable outcomes.

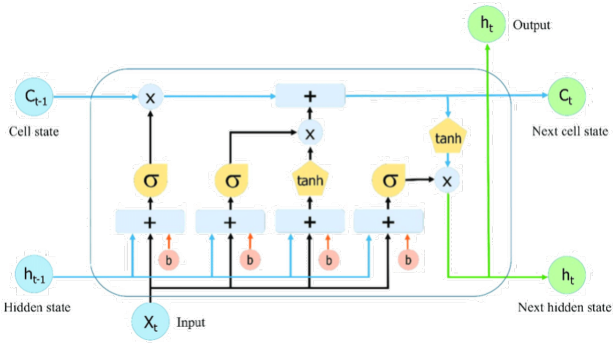

Fig. 12: LSTM Architecture [119].

As shown in Figure 12, an LSTM cell with input feature x_t receives input data x at time t, and an input gate i_t regulates the input data's flow into the cell. The forget gate f_t determines when to discard the contents of the cell's internal state, while the output gate o_t governs the flow of information to the output. Eq. 3-Eq. 8 summarizes the cell function:

$$i_t = \sigma(U_i x_t + W_i h_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(U_f x_t + W_f h_{t-1} + b_f) \tag{4}$$

$$o_t = \sigma(U_o x_t + W_o h_{t-1} + b_o) \tag{5}$$

$$g_t = \sigma(U_g x_t + W_g h_{t-1} + b_g) \tag{6}$$

$$c_t = g_t \odot i_t + f_t \odot c_{t-1} \tag{7}$$

$$h_t = o_t \odot tanh(c_t) \tag{8}$$

The logistic sigmoid function is denoted by σ, and the element-wise vector product operation is denoted by $\odot$. At any given time t, an LSTM architecture includes three gates: an input gate i_t, a forget gate f_t, and an output gate o_t, as well as a memory cell c_t and a hidden unit h_t. The initial values for c_0 and h_0 can be initialized to zero. LSTM model parameters consist of weight matrices U and W, as well as a bias vector b.

Tam et al. [120] developed LSTM based model to learn the text sequence and find the relation between words or phrases for sentiment classification. This model also improves the semantic information of tweets and enhances the learning model's efficiency. Another LSTM model [110] was developed, to overcome the limitations of RNNs, which can learn long sequences of data with time lags. The significant advantage of using the LSTM model is the recurrent units that allow long-range learning. A hidden state in augmented form is also included with non-linearity which permits updating the states, propagating it without any modification, or resetting, by employing simple learned gating functions. Drif and Hadjoudj [42] have proposed, two multi-level LSTM models, one based on user and content-specific features and the other one based on user, content, and sentiment features. Drif and Hadjoudj [42] conducted a case study on social media platforms to gain insights into sentiment intensity and the influence of social networking platforms on political protests. They built an LSTM model to analyze the effects of sentiment, user, and content on the dissemination of information, using the learning ability of the model to predict retweetability. Zhu et al. [43] proposed a sentiment index for the Chinese housing market by analyzing the sentiment expressed on social media regarding house prices. Imran et al. [44] presented a research study to analyze public reactions to the novel Coronavirus and the subsequent actions taken by different countries, from different cultures. They leveraged LSTM to estimate the sentiment polarity and trained their model using emotions extracted from tweets to achieve higher accuracy on their dataset.

Unlike LSTM, where information moves from backward to forward, Bi-LSTM allows information to flow in both directions through two hidden states which avoids the need for decay in future data inclusion. Figure 13 depicts an architecture of a Bi-LSTM model where the forward $\overrightarrow{h}$ and backward $\overleftarrow{h}$ sequences respectively are represented by the red and green arrows and the calculations are mentioned as in Eq. 9-Eq. 11.

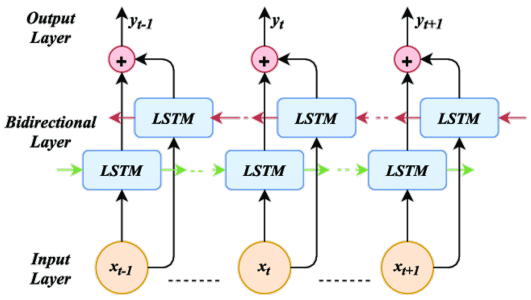

Fig. 13: Bi-LSTM Architecture [121].

$$\overrightarrow{h_t} = g(U_{\overrightarrow{h}} x_t + W_{\overrightarrow{h}} \overrightarrow{h_{t-1}} + b_{\overrightarrow{h}}) \tag{9}$$

$$\overleftarrow{h_t} = g(U_{\overleftarrow{h}} x_t + W_{\overleftarrow{h}} \overleftarrow{h_{t-1}} + b_{\overleftarrow{h}}) \quad (10)$$

$$y_t = g(V_{\overrightarrow{h}} \overrightarrow{h_t} + V_{\overleftarrow{h}} \overleftarrow{h_t} + b_y) \quad (11)$$

In the study presented by Schuster and Paliwal [122], a Bi-LSTM (Bi-directional long short-term memory) model is utilized which employs two independent recurrent networks to extract contextual relationships in both the forward and backward directions and enhances the limitations of the LSTM model in text sequence features. Feizollah et al. [37] proposed a sentiment classification method using CNNs, RNNs, and LSTM for tweets related to Halal cosmetics and Halal tourism., while Wang et al. [38] developed a sentiment prediction method using CNNs and Bidirectional LSTM to model multi-dimension and multi-level social media text to improve performance and textual semantic context. Blanco and Lourenço [123] have proposed an approach based on CNN and Bidirectional LSTM models for a better understanding of both optimistic and pessimistic sentiments related to COVID-19 discussions on Twitter. The authors have utilized a pre-trained transformer embedding for extracting significant semantic features from the data.

- **Gated Recurrent Unit (GRU)** is a frequently employed variation of the RNN model, which was introduced to overcome the challenge of the vanishing gradient problem [124]. This challenge is similar to the one addressed by the LSTM model. However, in various tasks, the GRU model has been found to outperform the LSTM model, with the exception of language modeling [110]. In contrast to the three gates in LSTM (*input*, *forget*, *output*), the GRU architecture (as depicted in Figure 14) is simpler comprising two gates: *update* (z_t) and *reset* (r_t), making its calculation less complex and effective in capturing long-term relationships between sequence elements [125].

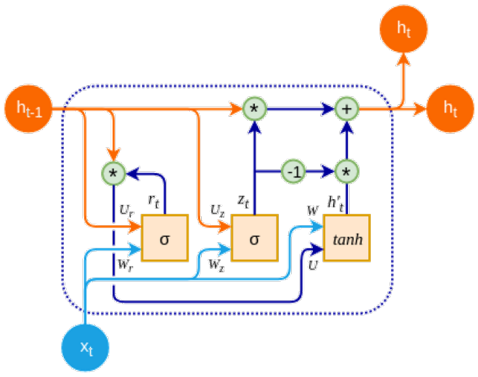

Fig. 14: GRU Architecture [124].

The reset gate is responsible for reducing the significance of the past hidden state (h_{t-1}) if it is deemed unnecessary for computing the new state, while the update gate determines the proportion of the previous state $(h_{(t-1)})$ that should be incorporated into the next state (h_t). The output state (h_t) is determined by a combination of the candidate output state $(\widetilde{h_t})$, the input vector (x_t), and the previous output state (h_{t-1}). The gates are updated using the sigmoid function (σ), and vector multiplication is accomplished through element-wise multiplication $(\odot)$. During training, the parameters for the gates (W_r, W_z, W_h) and biases (b_r, b_z, b_h) are learned. The calculation can be expressed using Eq. 12-Eq. 15

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (12)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (13)$$

$$\widetilde{h_t} = tanh(W x_t + U(r_t \times h_{t-1}) + b_h) \quad (14)$$

$$h_t = z_t \times h_{t-1} \oplus (1 - z_t) \times \widetilde{h_t} \quad (15)$$

Jabreel and Moreno [62] introduced a GRU-based model for the multi-emotion classification of Twitter data. This model was designed to be entirely data-driven and did not require external resources like emotion lexicons or POS taggers. However, the model faced difficulties when classifying emotions in extended and complicated text sequences. To improve the model's ability to capture intricate linguistic features and context, researchers have explored the use of an attention-based mechanism in sentiment analysis for Twitter.

- **Transformers-based** models, such as BERT (Bi-directional Encoder Representations from Transformers) [126, 127], RoBERTa (Robustly Optimized BERT pre-training approach) [128], XLNet (eXtreme MultiLingual Language Model) [129], and GPT (Generative Pre-trained Transformer) [130], has been fine-tuned for sentiment analysis on Twitter data. These models use attention mechanisms to weigh the importance of different words in a text and can identify key patterns and relationships between words, making them particularly effective for analyzing short and noisy text like Twitter posts.
The BERT model, introduced by Google AI in 2018 [111], is a bidirectional language model consisting of several encoders, an attention head, and a large feed-forward neural network. Each layer comprises a self-attention mechanism to process the input, which is then passed to the next layer via the feed-forward network. Initially, the input to the model is a sequence of words with tokens, and the output is a vector representation of the sequence. This representation, obtained from the first token of the input sequence, is used for sentiment classification. The output combined with a Softmax layer and feed-forward neural network is then used to determine the distribution of the target class. Figure 15 illustrates the BERT architecture for Twitter sentiment analysis.

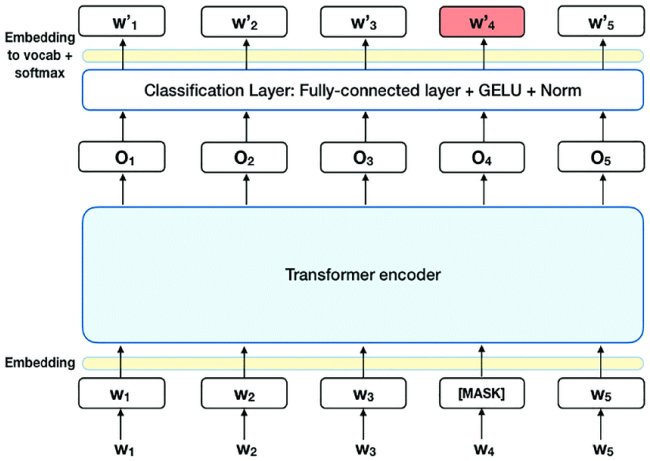

Fig. 15: BERT Architecture [131].

Bedi and Toshniwal [74] proposed BERT-based classification models for complaints and sentiment to improve the accuracy of energy-related tweets. Chandra and Saini [132] developed a framework to model the US general elections using two models, LSTM and BERT, to investigate if sentiment classification could predict election outcomes. Eke et al. [73] proposed a technique for sarcasm identification on IAC-v2 and Twitter data, using context-based features, employing three models including Bi-LSTM, BERT, and traditional machine learning.

Table V summarizes the literature on Twitter sentiment analysis.

TABLE V: DEEP LEARNING BASED MODELS

[Ref] Publication [Year]	Feature Extraction	Model	Attention	Domain
Non-Hybrid Models				
[64] Paredes-Valverde et al. [2017]	Word2Vec	CNN		Customer Feedback Analysis
[43] Zhu et al. [2018]	Word2Vec	LSTM		Real-Estate Analysis
[116] Zola et al. [2019]	POS, Word2Vec	MLP, CNN		Cross-domain Product Analysis
[62] Jabreel and Moreno [2019]	Word2Vec	GRU	✓	Opinion Mining
[67] Alharbi and de Doncker [2019]	Word2Vec	CNN, LSTM		Opinion Mining
[34] Yasir et al. [2020]	BOW	DNN		Interest-Rate Forecasting
[58] Mehta et al. [2021]	Standard Preprocessing	LSTM		Stock Market Prediction
[72] Pathak et al. [2021]	Word2Vec, TF-IDF	LSTM	✓	Topic Mining
[133] Basiri et al. [2021]	BOW, FastText	CNN, Bi-GRU, DistilBERT		Public Healthcare
[132] Chandra and Saini [2021]	Word2Vec	LSTM, BERT		Political View Mining
[42] Drif and Hadjoudj [2021]	User, Content & Sentiment features	LSTM		Opinion Mining
[73] Eke et al. [2021]	GloVe	Bi-LSTM, RNN, BERT	✓	Sarcasm Identification
[56] Demotte et al. [2021]	GloVe	Capsule Network		Tourism Planning
[134] Yang et al. [2021]	Word2Vec	LSTM	✓	Opinion Mining
[123] Blanco and Lourenço [2022]	GloVe	CNN, Bi-LSTM, BERT		Opinion Mining
[135] Alsayat [2022]	BOW, FastText	LSTM+Ensemble, BERT		Public Healthcare
[32] Ali et al. [2022]	Standard Preprocessing	DNN		Political View Mining
[74] Bedi and Toshniwal [2022]	Word2Vec, GloVe, FastText	BERT		Customer Feedback Analysis
[136] Wu et al. [2020]	Word2Vec	CNN		Opinion Mining
[44] Imran et al. [2020]	FastText, GloVe	DNN, LSTM, BERT, GRU, Bi-LSTM		Public Healthcare
[137] Chandrasekaran et al. [2022]	Visual Features	VGG-19, ResNet50V2, DenseNet-121		Opinion Mining
Hybrid Models				
[61] Ahmad et al. [2019]	Standard Preprocessing	LSTM+CNN		Crime Detection
[37] Feizollah et al. [2019]	Word2Seq, Word2Vec	CNN+Bi-RNN+BiLSTM		Tourism and Product Analysis
[138] Nguyen and Nguyen [2020]	LexW2Vs	LSTM, WAAN	✓	Opinion Mining
[139] Sadiq et al. [2020]	Visual Features	CNN+YOLO3+, CSPDarknet53		Disaster Management
[140] Visweswaran et al. [2020]	TF-IDF, GloVe, Word2Vec	CNN, LSTM, Bi-LSTM LSTM+CNN		Public Healthcare
[141] Salur and Aydin [2020]	Word2Vec, FastText, Character-level	LSTM, GRU, CNN, Bi-LSTM, CNN+Bi-LSTM		Market Strategies
[68] Alotaibi et al. [2021]	Standard Preprocessing	CNN, BiGRU, CNN+BiGRU+TrB	✓	Crime Prediction
[38] Wang et al. [2021]	Word2Vec	CNN, Bi-LSTM, CNN+BERT, CNN+BiLSTM, BERT+BiLSTM	✓	Opinion Mining
[120] Tam et al. [2021]	Word2Vec, GloVe	CNN+Bi-LSTM		Opinion Mining
[7] Lovera et al. [2021]	Standard Preprocessing	LSTM+Bi-LSTM, LIME		Opinion Mining
[60] Shehu et al. [2021]	Standard Preprocessing, Data Augmentation	Bi-GRU, CNN, HAN	✓	Opinion Mining
[66] Jain et al. [2021]	Standard Preprocessing	CNN+LSTM		Reviews Analysis
[65] Jalil et al. [2022]	Count Vectorizer, TF-IDF, GloVe, Word2Vec, FastText	DistilBERT, CNN+LSTM		Public Healthcare
[69] Abdalla and Özyurt [2021]	Word2Vec	CNN, Bi-LSTM, CNN+Bi-LSTM		Customer Feedback Analysis
[70] Umer et al. [2021]	TF-IDF, Word2Vec	CNN+LSTM		Hate Speech Detection
[142] Singh et al. [2022]	TF-IDF	LSTM+RNN	✓	Public Healthcare
[143] Galende et al. [2022]	TF	Bi-GRU		Crime Prediction
[59] Reshi et al. [2022]	TF-IDF	LSTM+GRU+RNN		Public Healthcare
[57] Swathi et al. [2022]	Standard Preprocessing	TLBO+LSTM		Stock Prediction

BOW: Bag-of-Words; **POS**: Parts-of-Speech; **TF**: Term Frequency; **TF-IDF**: Term Frequency-Inverse Document Frequency; **LexW2Vs**: Lexicon Embeddings; **BERT**: Bidirectional Encoder Representations from Transformers; **MLP**: Deep Multilayer Perceptron; **CNN**: Convolutional Neural Network; **DNN**: Deep Neural Network; **RNN**: Recurrent Neural Network; **TrB**: Transformer Block; **LSTM**: Long Short-Term Memory; **Bi-LSTM**: Bidirectional Long Short Term Memory; **GRU**: Gated Recurrent Unit; **Bi-GRU**: Bidirectional Gated Recurrent Unit; **WAAN**: Word Aspect Attention Network; **LIME**: Local Interpretable Model-Agnostic Explanations; **VGG**: Visual Geometry Group; **ResNet**: Residual Network; **DenseNet**: Dense Convolutional Network; **TLBO**: Teaching and Learning Based Optimization; **HAN**: Hierarchical Attention Network

C. Hybrid approaches

To combine the strengths of individual models, several researchers adopted hybrid models that integrate two or more approaches for instance lexicon and machine learning or lexicon and deep learning, and so on. The combination of these methods can help to overcome the limitations of each approach [144]. The advantage of combining learning-based approaches and lexicon is that it eliminates the need for manual labeling of training data as well as allows the measurement and detection of polarity at the conceptual level. Ngoge [145] developed a hybrid approach that combines ML techniques with lexicon methods to classify sentiment for the identification of terrorist activities. This approach uses SVM, Naïve Bayes classifier, and Maximum Entropy methods in combination with lexicon methods to predict patterns in tweets related to Kenya terrorist attacks. Gupta and Joshi [146] proposed a hybrid model that extracts feature vectors from SentiWordNet to build an SVM classifier for Twitter sentiment analysis. Du et al. [147] applied hierarchical ML to extract sentiment from opinions about HPV vaccines on Twitter and concluded the method to be highly efficient. Fadel and Cemil [27] presented a hybrid model to classify reviews on terrorist attacks posted on Twitter. The model utilizes a lexicon approach to generate a labeled training dataset and, an ML approach to finally build the model.

Although the combination of lexicon and machine learning methods has shown promising outcomes, there are still limitations that must be addressed to improve sentiment analysis efficiency. One such drawback is the reliance on the quality of the lexicon, which may not be adequate to handle complex semantic contexts, such as sarcasm, or filter out irrelevant words that add noise to reviews. In order to overcome the dependence on lexicon quality, various studies are based on hybrid deep-learning models that are capable of dealing with complex word patterns, thereby improving the performance of sentiment analysis tasks. Numerous hybrid deep learning models have been suggested in the literature to improve the performance of DL models used for Twitter sentiment analysis [142, 61]. Singh et al. [142] developed a hybrid DL model integrating LSTM and RNN models with attention layers to predict the sentiment of Twitter data related to COVID-19. Ahmad et al. [61] presented a joint approach of LSTM and CNN models to classify extremist-related tweets.

Salur and Aydin [141] proposed the amalgamation of various embeddings with multiple DL models, including LSTM, CNN, BiLSTM, and GRU, to extract features from word embeddings and then merge them for sentiment classification. Tam et al. [120] suggested a ConvBiLSTM model, which integrates Bi-LSTM and CNN to classify sentiment using Word2Vec and GloVe to obtain tweet embeddings. Shehu et al. [60] applied three data augmentation methods to increase the training size of stemmed Turkish Twitter data and subsequently used RNN, Hierarchical Attention Network (HAN), and CNN for sentiment analysis.

Jalil et al. [65] applied a hybrid model to analyze tweets collected on COVID-19 using various classifiers and feature sets. Jain et al. [66] suggested a hybrid CNN-LSTM model that uses word embedding to convert texts into vectors to classify sentiments of the text. Wu et al. [136] proposed a hybrid approach to summarize opinions on Chinese microblogging systems using CNN and the Ortony Clore Collins (OCC) model which is a rule-based export mechanism.

While the DL-hybrid and ML-hybrid models leverage the strengths of both deep learning and machine learning algorithms to achieve better sentiment analysis performance, they still have limitations in terms of capturing non-linear data complexity. Recently, new advancements have been made by exploring the field of Knowledge Graphs (KG), Graph Neural Networks (GNN), Capsule Networks (CN), etc., that we have detailed in the "other methods" subsection.

D. Other Methods

The "other methods" category comprises KG, GNN, and CN-based approaches that utilize the Twitter graph's properties and characteristics. While the GNNs are a subset of deep learning, their distinct characteristics in handling graph-structured data necessitate their separate classification. GNNs leverage the graph-based learning paradigm, which fundamentally differs from the standard feedforward learning approach used in most traditional deep learning models. In GNNs, each node in the graph is associated with a feature vector, and learning involves updating node representations by aggregating information from their neighboring nodes iteratively, where nodes represent users, hashtags, or words, and edges represent the complex, non-linear relationships among the nodes [148, 149]. Unlike traditional deep learning models such as Convolutional Neural Networks (CNNs) for images or Recurrent Neural Networks (RNNs) for sequences, which process fixed-sized inputs, GNNs operate directly on irregular graph structures. This recursive information propagation mechanism enables GNNs to capture complex patterns and dependencies within the data, especially in scenarios where traditional deep learning models struggle due to their fixed-sized input representations. Furthermore, unlike other approaches, these methods do not require large amounts of manually annotated data as they automatically collect annotated data using links between users and tweets, such as replies, followers, and previous tweets. However, they are domain-specific since the relationships and sentiment lexicon they use are tailored to the domain. These methods assume that sentiment and rating are interdependent and they address the issue of existing approaches where positive sentiment can be expressed using words with negative connotations.

Li et al. [30] proposed a graph-based technique, DWWP, which includes domain-specific word detection (DW) and word propagation (WP) methods. DW handles new words invented by users and converts the sentiment of words using Assembled Mutual Information (AMI), while WP includes semantic, and statistical similarity information, and manually calibrated sentiment scores that enhance the sentiment lexicon quality. Hussain and Cambria [31] analyzed knowledge-based reasoning using a vector space and support vector machine model, which utilizes lexical and graph representations for sentiment analysis. Aflakparast et al. [150] proposed a Bayesian graphical model to examine Twitter data. Demotte et al. [56] presented a Capsule network-based model that utilizes GloVe embeddings and dynamic/static routing to analyze social media content. Lovera et al. [7] developed a hybrid DL with a knowledge graph to analyze sentiment in a short text, such as Twitter posts. Aflakparast et al. [150] concluded that the results obtained from graph-based methods are promising, but there is still some ambiguity regarding the relationship between identified clusters and actual ratings. Moreover, these methods can be computationally demanding and time-consuming, and may not necessarily lead to improved accuracy.

Table VI presents a summary of the advantages (pros) and disadvantages (cons) of Twitter sentiment analysis methods based on different criteria.

IV. PUBLISHED DATA SOURCES AND TOOLS

This section details different versions of the Twitter dataset exploited in the existing literature and additional information about other similar datasets adopted for sentiment analysis. Furthermore, it details the employed performance metrics used for the evaluation of the proposed approaches [51]. Moreover, the section highlights the diverse tools and libraries leveraged for sentiment analysis tasks.

A. Dataset Description

In the presented work, we have broadly divided the various popular benchmark datasets used for sentiment analysis into two categories of *Twitter* and *other*. Table VII summarizes the dataset details and provides information about the size, polarity, source, and publications that have utilized them. Twitter has emerged as a prominent platform for sentiment analysis due to its large user base and the availability of real-time data. Researchers have used various versions of Twitter datasets for sentiment analysis, ranging from general datasets to domain-specific datasets. One commonly used dataset is the Sentiment140 dataset [151], which contains 1.6

million tweets labeled as positive or negative. It has been widely used for sentiment analysis research and benchmarking and has been utilized to evaluate the performance of various deep-learning models. Another common dataset is the SemEval, which contains tweets related to specific events or topics. The dataset provides labels for three categories: positive, negative, and neutral. *SemEval* [13] datasets have been used for sentiment analysis research and competition, providing a more challenging task due to the inclusion of neutral tweets. Several other domain-specific datasets include data related to politics, finance, and healthcare. These datasets provide a more targeted analysis of public sentiments within specific domains and can be useful for real-world applications.

It is a standard practice for researchers to use the Twitter dataset in sentiment analysis tasks. However, this data has its limitations such as short text (tweets have a word limit), noise, and the presence of sarcasm, irony, and slang, which can affect the accuracy of sentiment analysis models. In addition, the use of pre-labeled datasets may not always accurately represent the sentiment of the tweets, as the interpretation of sentiments can be very subjective.

To address these challenges, researchers have explored techniques for preprocessing and developing more robust sentiment analysis models that can handle noisy and ambiguous data. Additionally, researchers have explored the use of active and transfer learning techniques to improve the efficiency and accuracy of SA models with limited labeled data.

B. Evaluation Metrics

In addition to the gold standard evaluation metrics like *Precision* [32, 62, 74, 142], *Recall* [68, 73, 72], *F_score*[65, 69, 137, 143], *area under the curve (AUC)* [57] and *receiver operating characteristic curve (ROC)* [60], other performance metrics have been utilized in the literature to assess the performance of sentiment analysis models. This section outlines some of the commonly used evaluation metrics, along with their computation formulas. Table VIII presents an overview of various performance metrics adopted in literature to evaluate SA models.

- **Cohen's Kappa (CK)** is a measure of inter-annotator agreement that accounts for chance agreement [57, 152, 153]. It is defined as in Eq. 16:

$$CK = \frac{P_o - P_e}{1 - P_e} \tag{16}$$

  where P_o is the observed agreement, P_e is the expected agreement. P_o is calculated as the proportion of times the annotators agree, while P_e is calculated as the product of the marginal proportions of each label.
- **Mean Absolute Error (MAE)** is the average absolute difference between predicted and actual sentiment scores [154, 34]. It is defined as in Eq. 17:

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} \mid y_i - \widetilde{y_i} \mid \tag{17}$$

  where n is the number of instances, y_i is the actual sentiment score, and $\widetilde{y_i}$ is the predicted sentiment score for instance i respectively.
- **Root Mean Squared Error (RMSE)** is the square root of the average squared difference between predicted and actual sentiment scores [154, 155]. It is defined as in Eq. 18:

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (y_i - \widetilde{y_i})^2} \tag{18}$$

  where n is the number of instances, y_i is the actual sentiment score, and $\widetilde{y_i}$ is the predicted sentiment score for instance i respectively.
- **Spearman's Rank Correlation Coefficient (ρ)** measures the strength of the association between predicted and actual sentiment scores, taking into account the rank order of the scores rather than their absolute values [156]. It is defined as in Eq. 19:

$$\rho = 1 - \frac{6 \times \sum d_i^2}{n^2 \times (n-1)} \tag{19}$$

  where n is the number of instances, d_i is the difference between the rank of the predicted sentiment score and the rank of the actual sentiment score for i^{th} instance.

TABLE VI: PROS AND CONS OF VARIOUS TWITTER SENTIMENT ANALYSIS MODELS

Method($\longrightarrow$) / Criteria ($\downarrow$)	**Lexicon-based**	**Machine Learning-based**	**Deep Learning-based**
Reliance on Labeled Data	**Pros**: Not reliant on labeled data. **Cons**: Limited accuracy without additional data.	**Pros**: Requires labeled data **Cons**: The quality of results heavily depends on the quality and size of labeled data, and manual labeling can be time-consuming and costly.	**Pros**: Require less labeled data than traditional machine learning methods. **Cons**: May be affected by class imbalance and require labeled data for rare cases.
Language Agnostic	**Pros**: Works for any language. **Cons**: May not capture language-specific nuances.	**Pros**: Can be applied to different languages with appropriate preprocessing and feature engineering. **Cons**: Performance can be affected by the quality of preprocessing and language-specific characteristics.	**Pros**: Can be used for any language with enough labeled data. **Cons**: The model needs to be trained in a specific language and may not generalize well to other languages.
Accuracy	**Pros**: High accuracy for polarity classification. **Cons**: May not handle sarcasm, irony, or idiomatic expressions.	**Pros**: May be affected by class imbalance and require labeled data for rare cases. **Cons**: Accuracy can plateau due to limitations in the extracted features from data.	**Pros**: Can achieve state-of-the-art performance on sentiment analysis tasks. **Cons**: errors may still exist, especially in rare or ambiguous cases, and may not be robust to adversarial attacks.
Handling of Ambiguity	**Pros**: Able to handle ambiguous words and phrases. **Cons**: May not handle context-dependent words and phrases.	**Pros**: Can handle ambiguity well with the use of probabilistic models. **Cons**: Can be affected by the type and degree of ambiguity present in the data.	**Pros**: Can handle ambiguity and subtle nuances in language through the use of dense vector representations. **Cons**: May struggle with sarcasm, irony, and other forms of figurative language.
Ability to Capture Complex Relationships	**Pros**: Can capture complex word relationships such as synonyms, antonyms, and intensifiers. **Cons**: May not capture nuances of non-literal language such as metaphor and sarcasm.	**Pros**: Can capture complex relationships between words and the sentiment of the text. **Cons**: Performance can be limited by the complexity of the relationships present in the data.	**Pros**: Can learn complex, non-linear relationships between features and sentiment labels. **Cons**: May struggle with capturing causality, temporal relationships, and other complex language features.
Better Domain Adaptability	**Pros**: Can be adapted to specific domains. **Cons**: Limited coverage of uncommon words and phrases.	**Pros**: Can be adapted to different domains with the use of transfer learning techniques. **Cons**: Performance can be affected by the quality of the transfer learning and the similarity between the source and target domains.	**Pros**: Can be fine-tuned for specific domains with additional labeled data and achieve high performance. **Cons**: Performance may degrade significantly when moving to out-of-domain tasks.
Scalability	**Pros**: Fast and computationally inexpensive. **Cons**: Limited coverage of uncommon words and phrases.	**Pros**: Can scale to large datasets and be applied to real-time data with the use of distributed computing. **Cons**: Scalability can be limited by hardware and computational resources.	**Pros**: Can handle large amounts of data and be trained on distributed systems for faster training. **Cons**: Require significant computational resources and may not be accessible for researchers or organizations with limited resources.
End-to-End Learning	Not Applicable	**Pros**: Can perform end-to-end learning, where features are learned automatically from raw data, eliminating the need for manual feature engineering. **Cons**: Performance can be limited by the quality and size of the data used for training.	**Pros**: Can learn both feature representations and classification tasks in a single end-to-end pipeline. **Cons**: May not be suitable for applications where interpretability is important, such as legal or medical contexts.

TABLE VI: PROS AND CONS OF VARIOUS TWITTER SENTIMENT ANALYSIS MODELS

Transfer Learning	Not Applicable	**Pros**: Can use transfer learning to improve performance on new datasets with limited labeled data. **Cons**: Performance can be affected by the quality of the transfer learning and the similarity between the source and target domains.	**Pros**: Can leverage pre-trained models to achieve high performance on small datasets and out-of-domain tasks. **Cons**: May require some labeled data and may not always transfer well to tasks that are significantly different from the pre-training task.
Generalizability	**Pros**: May generalize to similar data. **Cons**: May not generalize to other datasets or domains.	**Pros**: Can generalize well to new and unseen data. **Cons**: Generalizability can be affected by the quality and size of the data used for training.	**Pros**: Can achieve high performance across different domains and datasets. **Cons**: Performance may degrade when moving to rare or specific sub-domains or datasets.
Limited Vocabulary Coverage	**Pros**: Limited coverage of uncommon words and phrases. **Cons**: May not capture nuanced differences between similar words.	**Pros**: Can be affected by the limited coverage of the vocabulary in the training data, which can result in errors for words outside the vocabulary. **Cons**: Limited by the availability and quality of domain-specific training data and the time required to update the lexicon.	**Pros**: Can handle out-of-vocabulary words through the use of character-based models and subword representations. **Cons**: May struggle with rare or infrequent words and require significant amounts of data for subword learning.
Difficulty Handling Irony and Sarcasm	**Cons**: May not identify and classify ironic or sarcastic statements.	**Pros**: Performance can be improved with the use of specialized lexicons and models designed for irony and sarcasm detection. **Cons**: May have difficulty handling irony and sarcasm due to the complex nature of these phenomena.	**Pros**: Can leverage context and long-term dependencies to better handle figurative language **Cons**: May struggle with subtle and complex forms of irony and sarcasm.
High Context Sensitivity	**Cons**: May struggle to identify sentiment in context-dependent statements.	**Pros**: Performance can be improved with the use of contextual information. **Cons**: May be affected by the context in which words are used and the social dynamics of Twitter.	**Pros**: Highly context sensitive. **Cons**: Limited by the availability and quality of contextual information.
Inability to Capture Negation	**Cons**: May not identify and handle negation well .	**Cons**: May struggle to accurately capture negation and its impact on sentiment.	**Pros**: Performance can be improved with the additional processing or training to capture negation. **Cons**: May misclassify sentiment in negated text.
Overfitting	Not Applicable	**Pros**: Performance can be improved with careful regularization and hyperparameter tuning to prevent overfitting. **Cons**: Can be prone to overfitting when model complexity is not properly constrained.	**Pros**: Can fit to noise or spurious correlations in the training data. **Cons**: Poor generalization to unseen data.
Lack of Transparency	**Pros**: Models are comparatively more interpretable than ML and DL-based models. **Cons**: May be difficult to understand how sentiment scores are assigned.	**Cons**: Limited interpretability and transparency of the model.	**Cons**: Difficult to interpret and understand the internal workings of the model.
Computation Intensive	**Pros**: Fast and computationally inexpensive. **Cons**: Limited coverage of uncommon words and phrases.	**Cons**: Requires significant computational resources to train and optimize.	**Cons**: Costly to scale up or deploy.

- **Kendall's Tau (τ)** is a measure of the strength of the association between predicted and actual sentiment scores, taking into account the number of pairwise disagreements [157]. It is defined as in Eq. 20:

$$\tau = \frac{n_c - n_d}{n \times (n-1)/2} \tag{20}$$

where n is the number of instances, n_c is the number of concordant pairs, n_d is the number of discordant pairs.

- **Hamming Loss (HL)** measures the fraction of labels that are incorrectly predicted for a given set of instances [57, 158]. It is defined as in Eq. 21:

$$HL = \frac{1}{n} \times \sum_{i=1}^{n} L(h(x_i), y_i) \tag{21}$$

where n is the number of instances, $h(x_i)$ predicted labels for instance i, y_i true labels for instance i, and L loss function, which is typically defined as the number of labels that are different between the predicted and true labels. In sentiment analysis, each instance (e.g., a tweet or a review) can be associated with multiple sentiment labels, such as positive, negative, neutral, or a combination of these. Hamming loss is used to evaluate the accuracy of a multi-label classifier in predicting the correct sentiment labels for each instance.

- **Jaccard Index (JI)** is also known as the *Jaccard similarity coefficient* or *Jaccard similarity index*) is an evaluation metric used in sentiment analysis and other natural language processing tasks to measure the similarity between two sets of labels [62]. It is defined as in Eq. 22:

$$JI = \frac{\mid A \cap B \mid}{\mid A \cup B \mid} \tag{22}$$

where A is the set of labels assigned by the model, B is the set of true labels, $\mid A \cap B \mid$ and $\mid A \cup B \mid$ represent the size of the intersection and the union between A and B respectively. In sentiment analysis, the Jaccard Index is used to evaluate the overlap between the predicted sentiment labels and the true sentiment labels for each instance (e.g., a tweet or a review). Jaccard Index's higher values indicate better performance.

- **Matthews Correlation Coefficient (MCC)** is an evaluation metric used in sentiment analysis and other classification tasks to measure the quality of the predictions made by a model. It takes into account true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [57, 159]. It is defined as in Eq. 23:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(P \times Q \times R \times S)}} \tag{23}$$

where, P, Q, R, and S denote $(TP+FP)$, $(TP+FN)$, $(TN+FP)$, and $(TN+FN)$ respectively. MCC ranges from -1 (total disagreement between the predictions and the true labels) to 1 (perfect agreement between the predictions and the true labels), where higher values indicate better performance. In sentiment analysis, MCC is used to evaluate the overall performance of a binary classifier in predicting the correct sentiment label for each instance (e.g., a tweet or a review).

- **Logarithmic Loss (Log Loss)** is used in binary classification tasks to measure the performance of a probabilistic classifier in predicting the correct label for each instance. It calculates the difference between the predicted probabilities and the true binary labels and penalizes high-confidence wrong predictions more than low-confidence ones [57]. It is defined as in Eq. 24:

$$LogLoss = \frac{1}{n} \times \sum (y \times log(p) + (1-y) \times log(1-p)) \tag{24}$$

where n is the total number of instances, y is the true binary label (0 or 1), and p is the predicted probability of the positive class (i.e., the sentiment label "positive").

C. Resources and Tools for Sentiment Analysis

Over the past few years, sentiment analysis on Twitter using deep learning has gained substantial attention. This has led to the development of various software tools and libraries that can be used to implement and evaluate these models. In this section, the commonly used software tools and libraries in state-of-the-art are compiled as presented in Table IX. The majority of the implementations use Python 3.x, along with popular deep learning libraries such as PyTorch, Keras, and TensorFlow. Additionally, some implementations also use the MATLAB platform.

- **NLTK** (*Natural Language Toolkit*) is a Python library that offers a comprehensive suite of tools and resources for natural language processing, including functions for tokenization, stemming, sentiment analysis, and text classification.
- **Scikit-learn** is a Python-based machine-learning library that offers a wide range of supervised and unsupervised learning algorithms. It provides tools for text classification, sentiment analysis, and feature extraction.
- **TensorFlow** is an open-source machine learning library created by Google that provides a wide range of tools for building and training deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers.
- **PyTorch** is an open-source machine learning library developed by Facebook. It offers tools for building and training deep learning models, including CNNs, RNNs, and transformers. It has gained popularity due to its user-friendliness and flexibility.
- **Keras:** Keras is a high-level deep learning library that provides a user-friendly API for building and training deep learning models. It is built on top of TensorFlow and simplifies the process of building complex models.
- **Gensim** is a Python library used for topic modeling and natural language processing. It provides tools for text preprocessing, topic modeling, and similarity calculation.
- **Word2vec** is a well-known algorithm used for word embeddings, generating dense vector representations of words that can be utilized as input to deep learning models for sentiment analysis.

TABLE VII: SUMMARY OF DATASETS

[Ref] Publication [Year]	Dataset		#Tweets	Polarity	Data Source
	Twitter	Others			
[64] Paredes-Valverde et al. [2017]	✓		100k	Pos/Neg	Dataset available on request.
[43] Zhu et al. [2018]		✓	1.2MM	Pos/Neg	Dataset available on request.
[155] Mahendhiran and Kannimuthu [2018]		✓	47 videos	Pos/Neg/Neu	https://dl.acm.org/doi/pdf/10.1145/2070481.2070509
[67] Alharbi and de Doncker [2019]	✓		∼3.7k+ ∼1.12k	Pos/Neg	https://aclanthology.org/S16-1081 https://aclanthology.org/S16-1082
[37] Feizollah et al. [2019]	✓	✓	∼84k	Pos/Neg	Dataset available on request.
[116] Zola et al. [2019]	✓	✓	∼75k+∼75k+∼5.8k	Pos/Neg/Neu	https://github.com/paolazola/Cross-source-crossdomain-sentiment-analysis
[62] Jabreel and Moreno [2019]	✓		∼10k	Multi-class	https://competitions.codalab.org/competitions/17751#learn_ the_details-datasets
[61] Ahmad et al. [2019]	✓		∼21k	Extremist/ Non-Extremist	Dataset available on request.
[136] Wu et al. [2020]		✓	∼23k+∼4.5k+∼16k	Pos/Neg/Neu	Dataset available on request.
[44] Imran et al. [2020]	✓		∼0.46MM	Pos/Neg	https://www.kaggle.com/datasets/smid80/ coronavirus-covid19-tweets
[139] Sadiq et al. [2020]	✓		1MM images	Pos/Neg/Neu	Dataset available on request.
[140] Visweswaran et al. [2020]	✓		∼11.5k	Pos/Neg	Dataset available on request.
[138] Nguyen and Nguyen [2020]	✓	✓	∼3k+∼4.2k+∼7k	Pos/Neg/Neu	https://aclanthology.org/S14-2004/
[141] Salur and Aydin [2020]	✓		∼1.7k	Pos/Neg/Neu	Dataset available on request.
[120] Tam et al. [2021	✓	✓	∼0.8MM	Pos/Neg	https://www.kaggle.com/atulanandjha/ stanford-sentimenttreebank-v2-sst2
[56] Demotte et al. [2021]	✓		∼2k+∼14.6k	Crowd:Pos/Neg/Neu Stanford:Pos/Neg	https://www.kaggle.com/datasets/crowdflower/ twitter-airline-sentiment https://www.kaggle.com/datasets/divyansh22/stsgold-dataset
[66] Jain et al. [2021]	✓		∼1.5k	Pos/Neg	www.data.world/crowdflower www.airlinequality.com
[38] Wang et al. [2021]	✓	✓	∼1.2MM+∼3.05MM+ ∼11k++∼18k	Pos/Neg	https://github.com/SophonPlus/ChineseNlpCorpus/tree/master/ datasets/weibo_senti_100k https://pan.baidu.com/s/16c93E5x373nsGozyWevITg?_at_= 1617119005896#list/path=%2F
[132] Chandra and Saini [2021]	✓		1.17MM	Pos/Neg/Neu	https://github.com/sydney-machine-learning/ sentimentanalysis-USelections
[42] Drif and Hadjoudj [2021]	✓		∼11.5k+18k	Multi-class	https://sourceforge.net/projects/alg-dialect-sentiment-dataset/
[7] Lovera et al. [2021	✓		16MM	Pos/Neg	http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
[60] Shehu et al. [2021]	✓		3k+10.5k	Pos/Neg/Neu	Dataset available on request.
[73] Eke et al. [2021]	✓	✓	∼2k+∼55k+∼6.5k	Pos/Neg	Dataset available on request.
[70] Umer et al. [2021]	✓		∼14.6k+∼23.5k+ ∼29.5k	Pos/Neg/Neu	https://www.kaggle.com/crowdflower/twitter-airline-sentiment www.kaggle.com/ramitsharma1994/ womens-ecommerce-clothing-review-prediction https://github.com/vedant-95/Twitter-Hate-Speech-Detection
[133] Basiri et al. [2021]	✓		∼1.6MM	Pos/Neg	Dataset available on request.
[72] Pathak et al. [2021]	✓		∼49k+ ∼46k+∼46k	Pos/Neg	https://data.mendeley.com/datasets/chx9mdyydb/1

TABLE VII: SUMMARY OF DATASETS

[134] Yang et al. [2021]	✓	✓	∼10k	Pos/Neg	https://alt.qcri.org/semeval2017/task4/index.php?id=data-and-tools
[58] Mehta et al. [2021]	✓		549+157+∼2.6k+∼10k+∼0.36MM	More Neg/Neg/Neu/Pos/More Pos	https://peerj.com/articles/cs-476/#supplementary-material
[68] Alotaibi et al.[2021]	✓		∼5.6k	Offensive/Non-Offensive	https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls https://github.com/zeeraktalat/hatespeech
[69] Abdalla and Özyurt [2021]	✓		50k+100k+200k	Pos/Neg	Dataset available on request.
[137] Chandrasekaran et al.[2022]	✓		1k images	Pos/Neg	https://data.world/crowdflower/image-sentiment-polarity
[143] Galende et al. [2022]	✓		4.5MM	Conspiracy/No Conspiracy	https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection
[59] Reshi et al. [2022]	✓		∼3.9k	Pos/Neg/Neu	Dataset available on request.
[57] Swathi et al. [2022]	✓		∼5.8k	Pos/Neg	Dataset available on request.
[74] Bedi and Toshniwal [2022]	✓		20k	Pos/Neg	Dataset available on request.
[142] Singh et al. [2022]	✓		170k	Pos/Neg	https://www.kaggle.com/gpreda/covid19-tweets
[32] Ali et al. [2022]	✓		∼3k	Pos/Neg/Neu	https://link.springer.com/article/10.1007/s00500-021-06569-5/tables/1
[123] Blanco and Lourenço [2022]	✓		∼0.15MM	Optimistic/Non-Optimistic/Pessimistic/Non-Pessimistic	https://www.sciencedirect.com/science/article/pii/S0306457322000437?via%3Dihub#sec0023
[135] Alsayat [2022]	✓	✓	18k	Pos/Neg	https://snap.stanford.edu/data/web-Amazon.html https://huggingface.co/datasets/yelp_review_full
[65]Jalil et al. [2022]	✓		90k+30k+ 30k+30k	Pos/Neg/Neu	https://github.com/usmaann/COVIDSenti
Pos: Positive; **Neg**: Negative; **Neu**: Neutral					

- **GloVe** (*Global Vectors for Word Representation*) is another widely used algorithm for word embeddings. It relies on co-occurrence statistics and produces dense vector representations of words that capture semantic relationships.

The earlier mentioned software tools and libraries provide a range of functionalities and resources for deep learning-based Twitter sentiment analysis. They are typically employed for preprocessing, feature extraction, model building, and evaluation. Nevertheless, selecting tools or libraries should be carefully considered based on the research question and the problem at hand.

TABLE VIII: SUMMARY OF COMMONLY USED PERFORMANCE METRICS IN SENTIMENT ANALYSIS

Metrics	Publications
Accuracy	[32, 37, 38, 116, 42, 132, 74, 120, 142, 61, 56, 136, 141, 43, 65, 66, 67, 68, 69, 70, 73, 72, 44, 139, 58, 59, 57, 137, 143, 155, 135, 138, 133, 134, 60]
Precision	[32, 37, 116, 62, 74, 142, 61, 120, 141, 7, 43, 64, 65, 66, 67, 68, 70, 73, 137, 143, 155, 140, 139, 58, 59, 57, 138, 133, 134]
Recall	[32, 37, 116, 62, 32, 142, 61, 56, 7, 120, 64, 141, 137, 143, 155, 139, 58, 59, 57, 138, 133, 65, 66, 67, 68, 69, 70, 73, 72]
F-Score	[32, 37, 116, 62, 132, 74, 142, 61, 56, 120, 7, 60, 141, 64, 65, 66, 67, 68, 69, 137, 143, 155, 44, 70, 139, 140, 73, 57, 59, 58, 72, 138, 133, 134]
RMSE	[34, 155]
MAE	[34]
AUC	[57, 60, 116, 140, 143]
RTM (Runtime)	[60]
Sensitivity	[57, 73, 116]
Specificity	[57, 73, 116]
Mean (Std.)	[132]
Jaccard Index	[62]
Kappa	[57, 155, 141, 140]
Hamming Loss	[57]
MCC	[57]
Log Loss	[57]

TABLE IX: SUMMARY OF TOOLS AND LIBRARIES

Tool	Publications	Link
Python	[37, 116, 132, 61, 7, 43, 68, 69, 44, 137, 140, 59, 133, 72]	https://www.python.org/
Scikit-learn	[59, 140]	https://scikit-learn.org/stable/
Tensorflow	[61, 136, 56, 7, 44, 64, 69, 73, 72, 59]	https://www.tensorflow.org/
PyTorch	[123]	https://pytorch.org/
Keras	[37, 116, 42, 61, 120, 141, 7, 66, 68, 69, 137, 44, 140, 73, 72]	https://keras.io/
Google Colab	[44, 68]	https://research.google.com/colaboratory
Theano	[43]	https://pypi.org/project/Theano/
Weka	[67]	https://www.weka.io/
Gensim	[29]	https://radimrehurek.com/gensim/
Word2vec	[64, 43, 116, 62, 67, 72, 132, 134, 37, 141, 38, 74, 120, 65]	https://pypi.org/project/word2vec/
GloVe	[73, 56, 123, 74, 44, 140, 120, 65]	https://pypi.org/project/glove/

V. PRACTICAL APPLICATIONS AND RELATED CASE STUDIES

Sentiment analysis has garnered significant attention from the research community due to its diverse use cases such as in social media, business, politics, healthcare, and tourism domains as illustrated in Figure 16. This section highlights the various practical business applications of sentiment analysis through related real-world case studies.

- **Brand Reputation Management:** Sentiment analysis can be used to monitor the online reputation of brands. A company can leverage it to monitor social media and determine how customers are reacting to their products or services. This information can help them improve their marketing strategies, create targeted campaigns, and ultimately increase sales. For instance, a study conducted by Zaki Ahmed and Rodríguez-Díaz [160] used sentiment analysis to monitor the online reputation of various airlines. The study analyzed tweets containing airline names and categorized the sentiments as positive, negative, or neutral. The results provided insights to airlines for reputation management and to improve their services. Another case study is the analysis of customer reviews for online food delivery apps to identify areas of improvement in their services [161].
- **Structure Marketing Strategies:** Sentiment analysis can be useful in formulating marketing strategies and marketing forecasting. A study by Lehrer et al. [162] suggests a deep learning-based technique to evaluate the polarity of sentiments on Twitter at an hourly rate. The proposed method considers mixed data sampling, resulting in a lower reduction of past data, which makes it highly appropriate for this novel source of data.
- **Political Opinion Mining:** Recently SA is increasingly used in politics to monitor public opinions and identify the sentiment behind political campaigns. It is useful for understanding and modeling voter behavior during political campaigns or activism, and can even indicate the outcome of an election. In a case study by Chandra and Saini [132], conducted on the US Presidential election, sentiment analysis was used to analyze Twitter data and identify the sentiment of the people towards the candidates. The proposed framework for modeling US general elections is based on LSTM and BERT models to predict voter sentiment. Another study conducted by Ali et al. [32] proposed a DNN model for sentiment analysis to predict 2018 general election results in Pakistan using Twitter opinions.
- **Customer Feedback Analysis:** In the realm of e-commerce and business intelligence, organizations can analyze customer feedback and reviews to understand the strengths and weaknesses of a business. Sentiment analysis is used to gain insights and opinions of users about products or events and to gain a deep understanding of customer interests and industry trends. Jain and Dandannavar [105] proposed a fast, scalable, and flexible sentiment analysis model on the Twitter dataset that uses Apache Spark and some machine learning models. Yasir et al. [34] deployed a DNN model to forecast the interest rate of five countries.
- **Finance Management:** Sentiment analysis can help investors make better decisions by providing insights into market sentiment. For example, investors can use sentiment analysis to analyze news articles and social media to understand the sentiment behind market movements. One such case study is the analysis of the sentiment of tweets related to the stock market and predicted changes in stock prices by Swathi et al. [57].
- **Public Healthcare:** Sentiment analysis can also be used to monitor public health. For instance, a study conducted by Reshi et al. [59] analyzed tweets related to COVID-19 and identified the areas where the outbreak was most severe. The results were used to improve public health policies.
- **Medical Services:** Healthcare providers can use sentiment analysis to analyze patient feedback and determine areas for improvement in their services. Opinion mining in health-related contexts is explored in [163], where the researcher offers new methods and a medical lexicon to assist patients and experts in explaining diseases and symptoms. The study used text processing and traditional machine learning methods as well.
- **Disaster Assessment, Response, and Management:** Sentiment analysis can be used to analyze social media data during dis-

asters to identify the areas affected, assess the public sentiment towards the disaster, and provide real-time updates to the public. A deep sentiment and activity analyzer combined with a deep human count tracker is proposed by Sadiq et al. [139] to track the number of people present in disaster-related visual content and analyze their sentiments.

- **Crime Prediction:** Sentiment analysis can be used for the identification and classification of potential criminal activities or terrorist groups. Ahmad et al. [61] proposed a tweet classification system using LSTM and CNN models to categorize tweets into extremist or non-extremist groups. Alotaibi et al. [68] developed an automatic cyberbullying approach for detecting aggressive behavior on Twitter by utilizing a bi-directional GRU, CNN model, and transformer block to catalog tweet sentiment as aggressive or not aggressive.
- **Tourism Planning:** Tourism is an important industry that is greatly influenced by public opinion. Sentiment analysis can be used to analyze user reviews and social media data to understand the satisfaction level of tourists and identify areas for improvement. Combining geo-location information with sentiment analysis can provide an effective plan for tourist destinations. Paolanti et al. [164] proposed a DNN approach for finding the sentiment of a widely known tourism venue, Cilento in Southern Italy.
- **Recommendation Systems:** Sentiment analysis can also benefit recommendation systems to offer personalized user recommendations. A hybrid CNN-LSTM model is suggested by Jain et al. [66] to classify the sentiment of customer reviews to further recommend user-personalized products. Preethi et al. [165] developed an RNN to analyze sentiments in reviews and improve movie and restaurant recommendations. Additionally, sentiment analysis can also aid in behavioral analysis in commodity markets [166].

VI. Research Gaps and Future Perspectives

Though deep learning models have shown significant evolution and excellent outcomes in the area of sentiment analysis, there exist several research gaps and open challenges which need further exploration. This section discusses the current research gaps and potential future directions for sentiment analysis research.

- **Decision-making Tool:** Deep learning models find their usage in various industries, including marketing, service, government, and academia, to analyze sentiment in decision-making problems. These models can be modified and adopted to achieve high accuracy, taking into account the complexities of textual analysis for practical applications. Numerous studies indicate that noisy features may negatively impact classification outcomes, hence DL methods can be designed to optimize features in an iterative process [7, 38, 120]. Additionally, the models can be improved to perform opinion mining, sentiment analysis, and topic detection simultaneously.
- **Processing Short Sequences:** Dealing with short sequences of social media text content that have varying content and background information is a challenging task. When it comes to processing such short sequences, dynamic routing is not as effective as static routing algorithms due to the variability of background details. However, this issue can be addressed by using Attention-based capsule networks [56] along with dynamic routing algorithms to extract relations for text content processing and sentiment analysis. Moreover, integrating contextual embedding with capsule-based models can lead to better performance as this technique has proven effective in most deep-learning approaches.
- **Handling Large Datasets:** One of the research gaps in SA on Twitter is the need to handle large datasets. Deep learning models have shown promising results in sentiment analysis on Twitter, but they require large datasets for effective optimization of the model parameters. The current state-of-the-art methods for sentiment analysis on Twitter, such as those presented in [68, 69], can be improved by applying these models to larger datasets. Therefore, future research could focus on developing new methods to handle large datasets that can improve the performance of sentiment analysis models on Twitter.
- **Handling Data Sparseness:** Handling data sparseness refers to the challenge of building effective sentiment analysis models when the dataset is limited or incomplete. Deep learning models have shown promising results in sentiment analysis tasks, but they require large datasets to perform well and optimize their parameters [116]. By doing so, it may be possible to address the issue of data sparsity and improve the accuracy and generalization ability of sentiment analysis models on Twitter.
- **Limited Attention to Domain-Specific SA:** The focus of sentiment analysis research has been predominantly on general sentiment analysis models, with limited attention paid to domain-specific sentiment analysis. There is a pressing need for models that can effectively analyze sentiment in specialized fields such as medical[163], financial[141], or legal data [68]. Future research should explore methods for developing domain-specific sentiment analysis models that can accurately capture the nuances of sentiment within these specific contexts.
- **Robustness and Reliability of Models:** The robustness and reliability of sentiment analysis models are an important area of research that needs to be addressed. The existing models are not robust enough to handle sarcasm, irony, and figurative language, which are prevalent in social media platforms, especially Twitter. The models often misinterpret these nuances and produce inaccurate results. Therefore, future research should focus on developing more robust models that can handle these language intricacies and improve the reliability of sentiment analysis results [64, 74]. Additionally, research can also explore the impact of linguistic and cultural differences on the accuracy of sentiment analysis models.
- **Interpretability of Models:** Interpretability of models refers to the ability to understand the reasoning behind a model's predictions. Most of the DL-based sentiment analysis models are considered black-box models because they operate on complex computations and are difficult to interpret [37, 141, 139]. This makes it challenging for users to trust the model's predictions and understand how they were generated. In recent years, there has been a growing interest in developing more transparent models, also known as *explainable AI*, that can provide insight into the reasoning behind their predictions. This approach could help increase trust in the model and improve its usefulness by allowing users to understand and potentially correct any biases or errors. Therefore, there is a need for research to develop more transparent sentiment analysis models that can provide explanations for their predictions.
- **Performance Measures:** Current evaluation metrics used in sentiment analysis research focus mainly on accuracy [32, 38]. There is a need for more comprehensive evaluation metrics that take into account the nuances of sentiment analysis, other than gold standard performance measures such as precision, recall, and F1-score [51].
- **Incorporating User Feedback:** Developing sentiment analysis models that can learn from user feedback by identifying user patterns and can adapt to dynamic user preferences may improve the accuracy of sentiment analysis models. Additionally, researchers can explore the use of interactive sentiment analysis tools that allow users to provide feedback in real-time, enabling the model to adapt to changing sentiments and preferences [60, 69].
- **Integrating Multiple Modalities:** Sentiment analysis has traditionally been limited to analyzing only textual data, but the incorporation of multiple modalities such as audio, video, and images can provide richer information for sentiment analysis [137]. Future research can focus on developing more advanced multimodal sentiment analysis models that can integrate mul-

Fig. 16: Twitter-based Sentiment Analysis Applications.

tiple modalities that could better capture the complexity and variability of human emotions and potentially provide more accurate and nuanced results in real-world settings.

VII. Conclusions

To sum up, the potential of using deep learning for sentiment analysis on Twitter has been widely recognized and has become an important research field due to the vast amount of user-generated content. The present work provides a comprehensive overview of the latest advances in deep learning techniques for sentiment analysis on Twitter. The current work outlines various preprocessing steps and word embeddings required for this task. The work presents a simplified taxonomy that bifurcates the literature into two categories: *conventional* (lexicon and machine-learning) and *deep learning* approaches, along with their respective pros and cons. Additionally, the paper summarizes various practical applications of sentiment analysis and identifies research gaps as well as domain-specific challenges. Furthermore, various metrics adopted by different studies to evaluate the models' performance are also reviewed in this work.

Overall, deep learning-based methods have shown great promise in sentiment analysis on Twitter, as they can capture complex language patterns and handle the noise and sparsity of data. Moreover, techniques like fine-tuning and transfer learning have proved effective in adapting pre-trained models to Twitter-specific datasets. However, there are still several challenges that need to be addressed in sentiment analysis on Twitter. One of the main challenges is dealing with the noise and complexity of Twitter data, such as short text, spelling mistakes, abbreviations, slang, and emojis. Another challenge is the imbalance of sentiments in data, with more instances of neutral or negative sentiments compared to positive sentiments. Also, the integration of sentiment analysis with other NLP techniques such as entity recognition and summarization could provide further insights and improve the performance of these models. In essence, this survey demonstrates that deep learning methods have advanced sentiment analysis on Twitter and that there is still much room for further improvement.

References

[1] N. Girdhar and K. K. Bharadwaj, "Signed social networks: a survey," in *Advances in Computing and Data Sciences: First International Conference, ICACDS 2016, Ghaziabad, India, November 11-12, 2016, Revised Selected Papers 1*. Springer, 2017, pp. 326–335.

[2] ——, "Community detection in signed social networks using multiobjective genetic algorithm," *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 788–804, 2019.

[3] N. Girdhar, S. Minz, and K. K. Bharadwaj, "Link prediction in signed social networks based on fuzzy computational model of trust and distrust," *Soft Computing*, vol. 23, pp. 12 123–12 138, 2019.

[4] N. Girdhar and K. K. Bharadwaj, "Social status computation for nodes of overlapping communities in directed signed social networks," *Integrated Intelligent Computing, Communication and Security*, pp. 49–57, 2019.

[5] ——, "Friends recommender system based on status (statusfrs) for users of overlapping communities in directed," *Applications of Artificial Intelligence Techniques in Engineering: SIGMA 2018, Volume 1*, vol. 698, p. 225, 2018.

[6] ——, "Mining of influencers in signed social networks: A memetic approach," in *Intelligent Human Computer Interaction: 10th International Conference, IHCI 2018, Allahabad, India, December 7–9, 2018, Proceedings 10*. Springer, 2018, pp. 306–316.

[7] F. A. Lovera, Y. C. Cardinale, and M. N. Homsi, "Sentiment analysis in twitter based on knowledge graph and deep learning classification," *Electronics*, vol. 10, no. 22, p. 2739, 2021.

[8] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *A practical guide to sentiment analysis*, pp. 1–10, 2017.

[9] E. Chu and D. Roy, "Audio-visual sentiment analysis for learning emotional arcs in movies," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 829–834.

[10] J. C. De Albornoz, L. Plaza, P. Gervás, and A. Díaz, "A joint model of feature mining and sentiment analysis for product review rating," in *Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings 33*. Springer, 2011, pp. 55–66.

[11] M. A. Mirtalaie, O. K. Hussain, E. Chang, and F. K. Hussain, "Sentiment analysis of specific product's features using product tree for application in new product development," in *Advances in Intelligent Networking and Collaborative Systems: The 9th International Conference on Intelligent Networking and Collaborative Systems (INCoS-2017)*. Springer, 2018, pp. 82–95.

[12] D. J. S. Oliveira, P. H. d. S. Bermejo, and P. A. dos Santos, "Can social media reveal the preferences of voters? a comparison between sentiment analysis and traditional opinion polls," *Journal of Information Technology & Politics*, vol. 14, no. 1, pp. 34–45, 2017.

[13] S. Rosenthal, N. Farra, and P. Nakov, "Semeval-2017 task 4: Sentiment analysis in twitter," *arXiv preprint arXiv:1912.00741*, 2019.

[14] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of twitter data for predicting stock market movements," in *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*. IEEE, 2016, pp. 1345–1350.

[15] T. Chalothorn and J. Ellman, "Using sentiwordnet and sentiment analysis for detecting radical content on web forums," 2012.

[16] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.

[17] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[18] V. Kharde, P. Sonawane *et al.*, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971*, 2016.

[19] M. A. Al-Khalisy and H. B. Jehlol, "Terrorist affiliations identifying through twitter social media analysis using data mining and web mapping techniques," *Journal of Engineering and Applied Sciences*, vol. 13, no. 17, pp. 7459–7464, 2018.

[20] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II 8*. Springer, 2016, pp. 22–39.

[21] M. Kumar, R. Bhatia, and D. Rattan, "A survey of web crawlers for information retrieval," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 6, p. e1218, 2017.

[22] S. Mansour, "Social media analysis of user's responses to terrorism using sentiment analysis and text mining," *Procedia Computer Science*, vol. 140, pp. 95–103, 2018.

[23] T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin, and B. Adini, "Twitter in the cross fire—the use of social media in the westgate mall terror attack in kenya," *PloS one*, vol. 9, no. 8, p. e104136, 2014.

[24] Y. Wei, L. Singh, and S. Martin, "Identification of extremism on twitter," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 1251–1255.

[25] S. A. Azizan and I. A. Aziz, "Terrorism detection based on sentiment analysis using machine learning," *Journal of Engineering and Applied Sciences*, vol. 12, no. 3, pp. 691–698, 2017.

[26] T. B. Mirani and S. Sasi, "Sentiment analysis of isis related tweets using absolute location," in *2016 international conference on computational science and computational intelligence (CSCI)*. IEEE, 2016, pp. 1140–1145.

[27] I. A. Fadel and Ö. Cemil, "A sentiment analysis model for terrorist attacks reviews on twitter," *Sakarya University Journal of Science*, vol. 24, no. 6, pp. 1294–1302, 2020.

[28] L. G. Smith, L. Wakeford, T. F. Cribbin, J. Barnett, and W. K. Hou, "Detecting psychological change through mobilizing interactions and changes in extremist linguistic style," *Computers in Human Behavior*, vol. 108, p. 106298, 2020.

[29] A. Aleroud, N. Abu-Alsheeh, and E. Al-Shawakfa, "A graph proximity feature augmentation approach for identifying accounts of terrorists on twitter," *Computers & Security*, vol. 99, p. 102056, 2020.

[30] W. Li, K. Guo, Y. Shi, L. Zhu, and Y. Zheng, "Dwwp: Domain-specific new words detection and word propagation system for sentiment analysis in the tourism domain," *Knowledge-Based Systems*, vol. 146, pp. 203–214, 2018.

[31] A. Hussain and E. Cambria, "Semi-supervised learning for big social data analysis," *Neurocomputing*, vol. 275, pp. 1662–1673, 2018.

[32] H. Ali, H. Farman, H. Yar, Z. Khan, S. Habib, and A. Ammar, "Deep learning-based election results prediction using twitter activity," *Soft Computing*, vol. 26, no. 16, pp. 7535–7543, 2022.

[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[34] M. Yasir, S. Afzal, K. Latif, G. M. Chaudhary, N. Y. Malik, F. Shahzad, and O.-y. Song, "An efficient deep learning based model to predict interest rate using twitter sentiment," *Sustainability*, vol. 12, no. 4, p. 1660, 2020.

[35] F.-Y. Zhou, L.-P. Jin, J. Dong *et al.*, "Review of convolutional neural network," 2017.

[36] Y. Li and H. Dong, "Text emotion analysis based on cnn and bilstm network feature fusion," *Comput. Appl*, vol. 38, no. 11, pp. 29–34, 2018.

[37] A. Feizollah, S. Ainin, N. B. Anuar, N. A. B. Abdullah, and M. Hazim, "Halal products on twitter: Data extraction and sentiment analysis using stack of deep learning algorithms," *IEEE Access*, vol. 7, pp. 83 354–83 362, 2019.

[38] B. Wang, D. Shan, A. Fan, L. Liu, and J. Gao, "A sentiment classification method of web social media based on multidimensional and multilevel modeling," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1240–1249, 2021.

[39] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on bilstm," *Ieee Access*, vol. 7, pp. 51 522–51 532, 2019.

[40] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[41] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[42] A. Drif and K. Hadjoudj, "An opinion spread prediction model with twitter emotion analysis during algeria's hirak," *The Computer Journal*, vol. 64, no. 3, pp. 358–368, 2021.

[43] E. Zhu, J. Wu, H. Liu, and K. Li, "A sentiment index of the housing market: Text mining of narratives on social media," *Available at SSRN 3223566*, 2018.

[44] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra, "Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets," *Ieee Access*, vol. 8, pp. 181 074–181 090, 2020.

[45] A. Mittal and S. Patidar, "Sentiment analysis on twitter data: A survey," in *Proceedings of the 7th International Conference on Computer and Communications Management*, 2019, pp. 91–95.

[46] N. F. F. D. Silva, L. F. Coletta, and E. R. Hruschka, "A survey and comparative study of tweet sentiment analysis via semi-supervised learning," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, pp. 1–26, 2016.

[47] N. Azzouza, K. Akli-Astouati, A. Oussalah, and S. A. Bachir, "A real-time twitter sentiment analysis using an unsupervised method," in *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics*, 2017, pp. 1–10.

[48] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artificial Intelligence Review*, pp. 1–57, 2021.

[49] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[50] R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends and challenges," *ACM Computing Surveys*, 2023.

[51] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artificial Intelligence Review*, vol. 53, no. 6, pp. 4335–4385, 2020.

[52] S. Sharma and A. Jain, "Role of sentiment analysis in social media security and analytics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1366, 2020.

[53] S. Soni and A. Sharaff, "Sentiment analysis of customer reviews based on hidden markov model," in *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, 2015, pp. 1–5.

[54] P. N. Jain and A. S. Vaidya, "Analysis of social media based on terrorism—a review," *Vietnam Journal of Computer Science*, vol. 8, no. 01, pp. 1–21, 2021.

[55] N. Mehra, S. Khandelwal, and P. Patel, "Sentiment identification using maximum entropy analysis of movie reviews," *St anford Univer sity, USA in*, 2002.

[56] P. Demotte, K. Wijegunarathna, D. Meedeniya, and I. Perera, "Enhanced sentiment extraction architecture for social media content analysis using capsule networks," *Multimedia tools and applications*, pp. 1–26, 2021.

[57] T. Swathi, N. Kasiviswanath, and A. A. Rao, "An optimal deep learning-based lstm for stock price prediction using twitter sentiment analysis," *Applied Intelligence*, vol. 52, no. 12, pp. 13 675–13 688, 2022.

[58] P. Mehta, S. Pandya, and K. Kotecha, "Harvesting social media sentiment analysis to enhance stock market prediction using deep learning," *PeerJ Computer Science*, vol. 7, p. e476, 2021.

[59] A. A. Reshi, F. Rustam, W. Aljedaani, S. Shafi, A. Alhossan, Z. Alrabiah, A. Ahmad, H. Alsuwailem, T. A. Almangour, M. A. Alshammari *et al.*, "Covid-19 vaccination-related sentiments analysis: a case study using worldwide twitter dataset," in *Healthcare*, vol. 10, no. 3. MDPI, 2022, p. 411.

[60] H. A. Shehu, M. H. Sharif, M. H. U. Sharif, R. Datta, S. Tokat, S. Uyaver, H. Kusetogullari, and R. A. Ramadan, "Deep sentiment analysis: a case study on stemmed turkish twitter data," *IEEE Access*, vol. 9, pp. 56 836–56 854, 2021.

[61] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," *Human-centric Computing and Information Sciences*, vol. 9, pp. 1–23, 2019.

[62] M. Jabreel and A. Moreno, "A deep learning-based approach for multi-label emotion classification in tweets," *Applied Sciences*, vol. 9, no. 6, p. 1123, 2019.

[63] J. G. Harb, R. Ebeling, and K. Becker, "A framework to analyze the emotional reactions to mass violent events on twitter and influential factors," *Information Processing & Management*, vol. 57, no. 6, p. 102372, 2020.

[64] M. A. Paredes-Valverde, R. Colomo-Palacios, M. d. P. Salas-Zárate, and R. Valencia-García, "Sentiment analysis in spanish for improvement of products and services: A deep learning approach," *Scientific Programming*, vol. 2017, 2017.

[65] Z. Jalil, A. Abbasi, A. R. Javed, M. Badruddin Khan, M. H. Abul Hasanat, K. M. Malik, and A. K. J. Saudagar, "Covid-19 related sentiment analysis using state-of-the-art machine learning and deep learning techniques," *Frontiers in Public Health*, vol. 9, p. 2276, 2022.

[66] 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," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1–15, 2021.

[67] A. S. M. Alharbi and E. de Doncker, "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information," *Cognitive Systems Research*, vol. 54, pp. 50–61, 2019.

[68] M. Alotaibi, B. Alotaibi, and A. Razaque, "A multichannel deep learning framework for cyberbullying detection on social media," *Electronics*, vol. 10, no. 21, p. 2664, 2021.

[69] G. Abdalla and F. Özyurt, "Sentiment analysis of fast food companies with deep learning models," *The Computer Journal*, vol. 64, no. 3, pp. 383–390, 2021.

[70] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. Sang Choi, "Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model," *Computational Intelligence*, vol. 37, no. 1, pp. 409–434, 2021.

[71] C. P. D. Cyril, J. R. Beulah, N. Subramani, P. Mohan, A. Harshavardhan, and D. Sivabalaselvamani, "An automated learning model for sentiment analysis and data classification of twitter data using balanced ca-svm," *Concurrent Engineering*, vol. 29, no. 4, pp. 386–395, 2021.

[72] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," *Applied Soft Computing*, vol. 108, p. 107440, 2021.

[73] C. I. Eke, A. A. Norman, and L. Shuib, "Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and bert model," *IEEE Access*, vol. 9, pp. 48 501–48 518, 2021.

[74] J. Bedi and D. Toshniwal, "Citenergy: A bert based model to analyse citizens' energy-tweets," *Sustainable Cities and Society*, vol. 80, p. 103706, 2022.

[75] Y. Liu and M. Zhang, "Neural network methods for natural language processing," 2018.

[76] E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than bags of words: Sentiment analysis with word embeddings," *Communication Methods and Measures*, vol. 12, no. 2-3, pp. 140–157, 2018.

[77] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[78] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[79] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[80] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[81] T. Shi and Z. Liu, "Linking glove with word2vec," *arXiv preprint arXiv:1411.5595*, 2014.

[82] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[83] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[84] J. Xu and Q. Du, "A deep investigation into fasttext," in *2019 IEEE 21st international conference on high performance computing and communications; IEEE 17th international conference on smart city; IEEE 5th International conference on data science and systems (HPCC/SmartCity/DSS)*. IEEE, 2019, pp. 1714–1719.

[85] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations. corr abs/1802.05365 (2018)," *arXiv preprint arXiv:1802.05365*, 1802.

[86] F. Sebastiani and A. Esuli, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of the 5th international conference on language resources and evaluation*. European Language Resources Association (ELRA) Genoa, Italy, 2006, pp. 417–422.

[87] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[88] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 216–225.

[89] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[90] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Security Informatics*, vol. 4, no. 1, pp. 1–13, 2015.

[91] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table extraction using conditional random fields," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 235–242.

[92] M. R. Huq, A. Ahmad, and A. Rahman, "Sentiment analysis on twitter data using knn and svm," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017.

[93] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," *HP Laboratories, Technical Report HPL-2011*, vol. 89, pp. 1–8, 2011.

[94] P. Palanisamy, V. Yadav, and H. Elchuri, "Serendio: Simple and practical lexicon based approach to sentiment analysis," in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 543–548.

[95] B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain, "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach," *Cognitive Computation*, vol. 7, pp. 487–499, 2015.

[96] S. A. El Rahman, F. A. AlOtaibi, and W. A. AlShehri, "Sentiment analysis of twitter data," in *2019 international conference on computer and information sciences (ICCIS)*. IEEE, 2019, pp. 1–4.

[97] M. Mashuri *et al.*, "Sentiment analysis in twitter using lexicon based and polarity multiplication," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*. IEEE, 2019, pp. 365–368.

[98] S. Rani and J. Singh, "Sentiment analysis of tweets using support vector machine," *Int. J. Comput. Sci. Mob. Appl*, vol. 5, no. 10, pp. 83–91, 2017.

[99] A. L. Firmino Alves, C. D. S. Baptista, A. A. Firmino, M. G. d. Oliveira, and A. C. d. Paiva, "A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup," in *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, 2014, pp. 123–130.

[100] H. Wu, J. Li, and J. Xie, "Maximum entropy-based sentiment analysis of online product reviews in chinese," in *Automotive, Mechanical and Electrical Engineering*. CRC Press, 2017, pp. 559–562.

[101] E. Omer, "Using machine learning to identify jihadist messages on twitter," 2015.

[102] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, "Detecting multipliers of jihadism on twitter," in *2015 IEEE international conference on data mining workshop (ICDMW)*. IEEE, 2015, pp. 954–960.

[103] M. Nouh, J. R. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on twitter," in *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2019, pp. 98–103.

[104] A. Omar, T. M. Mahmoud, T. Abd-El-Hafeez, and A. Mahfouz, "Multi-label arabic text classification in online social networks," *Information Systems*, vol. 100, p. 101785, 2021.

[105] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016, pp. 628–632.

[106] M. Hartung, R. Klinger, F. Schmidtke, and L. Vogel, "Identifying right-wing extremism in german twitter profiles: A classification approach," in *Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings 22*. Springer, 2017, pp. 320–325.

[107] Z. U. Rehman, S. Abbas, M. A. Khan, G. Mustafa, H. Fayyaz, M. Hanif, and M. A. Saeed, "Understanding the language of isis: An empirical approach to detect radical content on twitter using machine learning," *Comput., Mater. Continua*, vol. 66, no. 2, pp. 1075–1090, 2021.

[108] W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali, M. Husnain, and G. S. Choi, "An empirical approach for extreme behavior identification through tweets using machine learning," *Applied Sciences*, vol. 9, no. 18, p. 3723, 2019.

[109] M. A. Masood and R. A. Abbasi, "Using graph embedding and machine learning to identify rebels on twitter," *Journal of Informetrics*, vol. 15, no. 1, p. 101121, 2021.

[110] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[111] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[112] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[113] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[114] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[115] J. Feng, X. He, Q. Teng, C. Ren, H. Chen, and Y. Li, "Reconstruction of porous media from extremely limited information using conditional generative adversarial networks," *Physical Review E*, vol. 100, no. 3, p. 033308, 2019.

[116] P. Zola, P. Cortez, C. Ragno, and E. Brentari, "Social media cross-source and cross-domain sentiment classification,"

International Journal of Information Technology & Decision Making, vol. 18, no. 05, pp. 1469–1499, 2019.

[117] H. Gu, Y. Wang, S. Hong, and G. Gui, "Blind channel identification aided generalized automatic modulation recognition based on deep learning," *IEEE Access*, vol. 7, pp. 110 722–110 729, 2019.

[118] W. Fang, J. Jiang, S. Lu, Y. Gong, Y. Tao, Y. Tang, P. Yan, H. Luo, and J. Liu, "A lstm algorithm estimating pseudo measurements for aiding ins during gnss signal outages," *Remote sensing*, vol. 12, no. 2, p. 256, 2020.

[119] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (lstm) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, 2019.

[120] S. Tam, R. B. Said, and Ö. Ö. Tanriöver, "A convbilstm deep learning model-based approach for twitter sentiment classification," *IEEE Access*, vol. 9, pp. 41 283–41 293, 2021.

[121] I. K. Ihianle, A. O. Nwajana, S. H. Ebenuwa, R. I. Otuka, K. Owa, and M. O. Orisatoki, "A deep learning approach for human activities recognition from multimodal sensing devices," *IEEE Access*, vol. 8, pp. 179 028–179 038, 2020.

[122] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[123] G. Blanco and A. Lourenço, "Optimism and pessimism analysis using deep learning on covid-19 related twitter conversations," *Information processing & management*, vol. 59, no. 3, p. 102918, 2022.

[124] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[125] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[126] S. Mann, J. Arora, M. Bhatia, R. Sharma, and R. Taragi, "Twitter sentiment analysis using enhanced bert," in *Intelligent Systems and Applications: Select Proceedings of ICISA 2022*. Springer, 2023, pp. 263–271.

[127] A. Bello, S.-C. Ng, and M.-F. Leung, "A bert framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, p. 506, 2023.

[128] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. T. Neves, "Unified benchmark and comparative evaluation for tweet classification," *Findings of the Association for Computational Linguistics*, 2020.

[129] Y. Cui, Y. Jiang, and H. Gu, "Novel sentiment analysis from twitter for stock change prediction," in *Data Mining and Big Data: 7th International Conference, DMBD 2022, Beijing, China, November 21–24, 2022, Proceedings, Part II*. Springer, 2023, pp. 160–172.

[130] M. U. Haque, I. Dharmadasa, Z. T. Sworna, R. N. Rajapakse, and H. Ahmad, "" i think this is the most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data," *arXiv preprint arXiv:2212.05856*, 2022.

[131] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics*, vol. 100, p. 100057, 2019.

[132] R. Chandra and R. Saini, "Biden vs trump: modeling us general elections using bert language model," *IEEE Access*, vol. 9, pp. 128 494–128 505, 2021.

[133] M. E. Basiri, S. Nemati, M. Abdar, S. Asadi, and U. R. Acharrya, "A novel fusion-based deep learning model for sentiment analysis of covid-19 tweets," *Knowledge-Based Systems*, vol. 228, p. 107242, 2021.

[134] J. Yang, X. Zou, W. Zhang, and H. Han, "Microblog sentiment analysis via embedding social contexts into an attentive lstm," *Engineering Applications of Artificial Intelligence*, vol. 97, p. 104048, 2021.

[135] A. Alsayat, "Improving sentiment analysis for social media applications using an ensemble deep learning language model," *Arabian Journal for Science and Engineering*, vol. 47, no. 2, pp. 2499–2511, 2022.

[136] P. Wu, X. Li, S. Shen, and D. He, "Social media opinion summarization using emotion cognition and convolutional neural networks," *International Journal of Information Management*, vol. 51, p. 101978, 2020.

[137] G. Chandrasekaran, N. Antoanela, G. Andrei, C. Monica, and J. Hemanth, "Visual sentiment analysis using deep learning models with social media data," *Applied Sciences*, vol. 12, no. 3, p. 1030, 2022.

[138] H.-T. Nguyen and L.-M. Nguyen, "Ilwaanet: an interactive lexicon-aware word-aspect attention network for aspect-level sentiment classification on social networking," *Expert Systems with Applications*, vol. 146, p. 113065, 2020.

[139] A. M. Sadiq, H. Ahn, and Y. B. Choi, "Human sentiment and activity recognition in disaster situations using social media images based on deep learning," *Sensors*, vol. 20, no. 24, p. 7115, 2020.

[140] S. Visweswaran, J. B. Colditz, P. O'Halloran, N.-R. Han, S. B. Taneja, J. Welling, K.-H. Chu, J. E. Sidani, and B. A. Primack, "Machine learning classifiers for twitter surveillance of vaping: comparative machine learning study," *Journal of Medical Internet Research*, vol. 22, no. 8, p. e17478, 2020.

[141] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58 080–58 093, 2020.

[142] C. Singh, T. Imam, S. Wibowo, and S. Grandhi, "A deep learning approach for sentiment analysis of covid-19 reviews," *Applied Sciences*, vol. 12, no. 8, p. 3709, 2022.

[143] B. A. Galende, G. Hernández-Peñaloza, S. Uribe, and F. Á. García, "Conspiracy or not? a deep learning approach to spot it on twitter," *IEEE Access*, vol. 10, pp. 38 370–38 378, 2022.

[144] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, 2020.

[145] L. A. Ngoge, "Real–time sentiment analysis for detection of terrorist activities in kenya," Ph.D. dissertation, Strathmore University, 2016.

[146] I. Gupta and N. Joshi, "Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic," *Journal of intelligent systems*, vol. 29, no. 1, pp. 1611–1625, 2020.

[147] J. Du, J. Xu, H.-Y. Song, and C. Tao, "Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data," *BMC medical informatics and decision making*, vol. 17, pp. 63–70, 2017.

[148] M. Wang and G. Hu, "A novel method for twitter sentiment analysis based on attentional-graph neural network," *Information*, vol. 11, no. 2, p. 92, 2020.

[149] W. Liao, B. Zeng, J. Liu, P. Wei, X. Cheng, and W. Zhang, "Multi-level graph neural network for text sentiment analysis," *Computers & Electrical Engineering*, vol. 92, p. 107096, 2021.

[150] M. Aflakparast, M. de Gunst, and W. van Wieringen, "Analysis of twitter data with the bayesian fused graphical lasso," *PloS one*, vol. 15, no. 7, p. e0235596, 2020.

[151] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[152] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[153] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *arXiv preprint cmp-lg/9602004*, 1996.

[154] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[155] P. Mahendhiran and S. Kannimuthu, "Deep learning techniques for polarity classification in multimodal sentiment analysis,"

International Journal of Information Technology & Decision Making, vol. 17, no. 03, pp. 883–910, 2018.

[156] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

[157] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[158] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.

[159] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[160] A. Zaki Ahmed and M. Rodríguez-Díaz, "Significant labels in sentiment analysis of online customer reviews of airlines," *Sustainability*, vol. 12, no. 20, p. 8683, 2020.

[161] S. K. Trivedi and A. Singh, "Twitter sentiment analysis of app based online food delivery companies," *Global Knowledge, Memory and Communication*, 2021.

[162] S. Lehrer, T. Xie, and T. Zeng, "Does high-frequency social media data improve forecasts of low-frequency consumer confidence measures?" *Journal of Financial Econometrics*, vol. 19, no. 5, pp. 910–933, 2021.

[163] R. Satapathy, E. Cambria, and A. Hussain, *Sentiment Analysis in the Bio-Medical Domain*. Springer, 2017.

[164] M. Paolanti, A. Mancini, E. Frontoni, A. Felicetti, L. Marinelli, E. Marcheggiani, and R. Pierdicca, "Tourism destination management using sentiment analysis and geo-location information: a deep learning approach," *Information Technology & Tourism*, vol. 23, pp. 241–264, 2021.

[165] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha, and S. Yenduri, "Application of deep learning to sentiment analysis for recommender system on cloud," in *2017 International conference on computer, information and telecommunication systems (CITS)*. IEEE, 2017, pp. 93–97.

[166] M. J. Keenan, *Advanced positioning, flow, and sentiment analysis in commodity markets: bridging fundamental and technical analysis*. John Wiley & Sons, 2020.