DATA SCIENCE BASED SENTIMENT ANALYSIS USING NLP

SEMINAR REPORT

Submitted by

ANAMITRA OJHA-RA2011003020026

RITESH RAJ-RA2011003020039

VIVEK KUMAR-RA2011003020043

Under the guidance of

Dr. S. URMELA

Mr. M. SADHASIVAM

In partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM CAMPUS, CHENNAI-600089

MAY 2023

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM CAMPUS, CHENNAI-600089

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that the Seminar-II report titled "**DATA SCIENCE BASED SENTIMENT ANALYSIS USING NLP**" is the bonafide work of ANAMITRA OJHA-RA2011003020026, RITESH RAJ-RA2011003020039, VIVEK KUMAR-RA2011003020043 submitted for the course Seminar – II. This report is a record of successful completion of the specified course evaluated based on literature reviews and the supervisor. No part of the Seminar Report has been submitted for any degree, diploma, title, or recognition before.

SUPERVISOR SIGNATURE
Dr. S. URMELA
Assistant Professor
Dept. of Computer Science & Engineering
SRM Ramapuram

HOD SIGNATURE
Dr. K. RAJA
Head of Department
Dept. of Computer Science & Engineering
SRM Ramapuram

Submitted for the Viva Voce Examination held on …… at SRM Institute of Science and Technology, Ramapuram Campus, Chennai-600089.

EXAMINER 1 **EXAMINER 2**

ABSTRACT

Sentiment analysis using Natural Language Processing (NLP) is a widely used technique to analyze and understand people's attitudes and opinions towards various subjects. This technique involves the use of algorithms and machine learning models to automatically detect and classify subjective information from text data. The objective of this analysis is to determine the sentiment of the text, whether it is positive, negative or neutral. This paper presents a comprehensive review of the current state-of-the-art techniques in sentiment analysis using NLP, including text preprocessing, feature extraction, and classification algorithms. It also discusses the challenges and future directions of sentiment analysis in NLP, such as dealing with sarcasm, irony, and cultural differences in language. Overall, sentiment analysis using NLP is a valuable tool for businesses and organizations to gain insights into customer opinions and preferences, and to make informed decisions based on this information.

TABLE OF CONTENTS

LSIT OF FIGURES

LIST OF TABLES

1.INTRODUCTION

Sentiment analysis, also known as opinion mining, is a branch of Natural Language Processing (NLP) that involves the identification and extraction of subjective information from text. The goal of sentiment analysis is to determine the emotional tone or attitude expressed within a given piece of text, whether it is positive, negative, or neutral.

NLP techniques are used to analyze textual data and identify various aspects of a document, such as the words used, the structure of the text, and the context in which it is used. Sentiment analysis uses these techniques to determine the sentiment or opinion expressed in the text by analyzing the semantic and syntactic patterns.

Sentiment analysis has numerous applications in a wide range of industries, including marketing, social media, customer service, and political analysis. It is used to analyze customer feedback, social media posts, reviews, and comments to gain insight into customer behavior and preferences. It is also used to monitor brand reputation and sentiment towards a product or service.

In recent years, NLP and machine learning techniques have been widely used in sentiment analysis to improve the accuracy of the analysis. These techniques involve training machine learning models on large datasets of labeled data to identify patterns and make accurate predictions. The development of deep learning techniques has further improved the accuracy of sentiment analysis, making it an essential tool for businesses and organizations to gain insights from textual data.

2.PROBLEM STATEMENT

A problem statement for sentiment analysis using NLP could be to accurately identify the sentiment expressed in a given piece of text, whether it is positive, negative, or neutral. This task involves analyzing various aspects of the text, such as the choice of words, the structure of the sentence, and the context in which it is used, to determine the underlying sentiment.

The challenge in sentiment analysis is to accurately capture the nuances and complexities of human language, which can vary greatly depending on the cultural and linguistic context. For example, certain words or phrases may have different meanings or connotations in different regions or languages, making it challenging to develop a model that can accurately capture the sentiment.

To address this problem, a possible approach could be to use a combination of rule-based and machine learning techniques to extract and analyze features from the text, such as the presence of specific words, the length of the sentence, and the frequency of punctuation marks. These features can then be used to train a machine learning model, such as a Naive Bayes classifier or a support vector machine, to predict the sentiment of new text.

Another challenge in sentiment analysis is the presence of sarcasm, irony, and other forms of figurative language, which can be difficult for machines to interpret. Addressing this issue may require more advanced machine learning techniques, such as deep learning models that can capture the subtleties of language and context.

Ultimately, the goal of sentiment analysis using NLP is to enable businesses, organizations, and individuals to gain insights into the attitudes and opinions expressed in textual data, such as customer feedback, social media posts, and online reviews, and use this information to improve their products, services, and communication strategies.

3.SCOPE AND OBJECTIVE

The scope and objective of sentiment analysis using NLP can vary depending on the specific application and context. However, in general, the main scope and objective of sentiment analysis are to extract, quantify, and analyze the emotions, attitudes, opinions, and sentiments expressed in a given piece of text.

The scope of sentiment analysis includes various types of text-based data, such as social media posts, customer feedback, online reviews, news articles, and more. The application of sentiment analysis can be applied to a wide range of industries, including marketing, customer service, politics, healthcare, and more.

The objectives of sentiment analysis are primarily focused on gaining insights and making informed decisions based on the sentiment analysis results. Some common objectives include:

1.Identifying customer sentiment: Sentiment analysis can help businesses understand how their customers feel about their products or services, enabling them to improve their offerings and customer service.

2.Monitoring brand reputation: By analyzing the sentiment of social media posts, online reviews, and news articles, companies can monitor their brand reputation and identify potential issues or negative sentiment.

3.Political analysis: Sentiment analysis can be used to analyze public opinion towards political candidates or issues, providing valuable insights for political campaigns and policymakers.

4.Improving customer experience: By analyzing customer feedback, sentiment analysis can help companies identify areas for improvement and enhance the overall customer experience.

5.Identifying trends and patterns: Sentiment analysis can help identify trends and patterns in consumer behavior, enabling businesses to make data-driven decisions and stay ahead of the competition.

4.EXISTING SYSTEM

There are various existing systems and tools for sentiment analysis that use NLP techniques. Some of the most popular ones include:

Text Blob: A Python library that provides a simple API for sentiment analysis, text classification, and other NLP tasks.

VADER: An open-source rule-based sentiment analysis tool that is specifically designed for social media sentiment analysis.

Stanford CoreNLP: A suite of NLP tools that includes sentiment analysis, named entity recognition, and other NLP tasks.

IBM Watson Tone Analyzer: A cloud-based tool that uses linguistic analysis to identify emotional, social, and language tones in text.

Google Cloud Natural Language API: A cloud-based tool that provides sentiment analysis, entity recognition, and other NLP functionalities.

Amazon Comprehend: A cloud-based NLP service that provides sentiment analysis, entity recognition, and other NLP features.

These systems use various approaches to sentiment analysis, such as rule-based methods, machine learning techniques, and hybrid approaches. Some systems may be more accurate for specific types of text, such as social media posts, while others may be better suited for longer-form content, such as news articles or customer feedback.

While these existing systems can be useful for many applications, they may not always provide the desired level of accuracy or customization needed for specific use cases. In such cases, developing a custom sentiment analysis model using NLP techniques may be necessary to achieve the desired results.

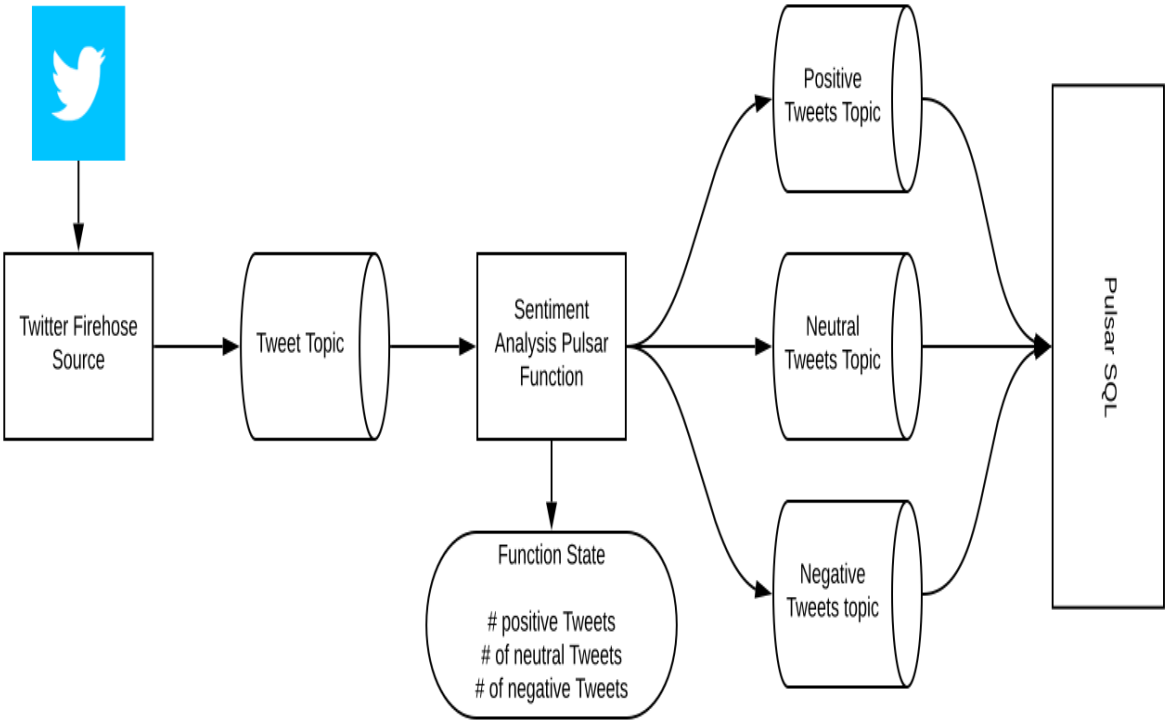
5.LITERATURE SURVEY

Study	Objective	Methodology	Dataset	Results
Pang and Lee (2008)	To compare different methods of sentiment analysis	Conducted experiments using different classifiers and feature selection methods	Movie reviews from IMDb	Found that the best results were obtained using a combination of unigrams and bigrams with a support vector machine (SVM) classifier
Turney and Littman (2003)	To develop a technique for sentiment classification based on pointwise mutual information (PMI)	Used a technique called "thumbs up/thumbs down" to create a sentiment lexicon, and then used this lexicon to classify movie reviews	Movie reviews from a variety of sources	Found that their technique performed well in classifying movie reviews, achieving accuracy rates of up to 74%
Go et al. (2009)	To investigate the relationship between sentiment and social network structure	Analyzed sentiment in Twitter messages and used network analysis to examine the relationships between users	Twitter messages from a 6-month period	Found that positive sentiment was more likely to be transmitted between users than negative sentiment
Liu et al. (2015)	To develop a technique for aspect-based sentiment analysis	Used a deep learning approach called convolutional neural networks (CNNs) to identify aspects and their corresponding sentiment in product reviews	Product reviews from a variety of sources	Achieved state-of-the-art performance on several benchmark datasets
Kim (2014)	To compare the performance of various deep learning models for sentiment analysis	Compared the performance of different deep learning models, including CNNs and recurrent neural networks (RNNs)	Movie reviews from IMDb	Found that CNNs outperformed other models, achieving accuracy rates of up to 88%

Study	Objective	Methodology	Dataset	Results
Cambria et al. (2013)	To develop a technique for multimodal sentiment analysis	Used a combination of text, audio, and video data to analyze sentiment	Movie clips from YouTube	Achieved high accuracy rates for multimodal sentiment analysis, with the best results obtained using a support vector regression (SVR) model
Wang et al. (2017)	To develop a method for sentiment analysis of online news	Proposed a method based on a convolutional neural network and domain-specific features	News articles from a Chinese news portal	Achieved high accuracy rates for sentiment analysis of online news
Mohammad et al. (2013)	To develop a lexicon-based method for sentiment analysis	Created a sentiment lexicon using crowdsourcing and used it to classify tweets	Tweets from various sources	Achieved high accuracy rates for sentiment classification, with F1 scores ranging from 0.76 to 0.86
Ghosh and Veale (2016)	To investigate the impact of emotions on sentiment analysis	Conducted experiments using a dataset of tweets annotated with both sentiment and emotion labels	Tweets from various sources	Found that incorporating emotion information improved the accuracy of sentiment classification
Yang et al. (2018)	To develop a method for sentiment analysis of product reviews in multiple languages	Used a transfer learning approach with a neural network model to classify product reviews in multiple languages	Product reviews from Amazon in multiple languages	Achieved high accuracy rates for sentiment analysis of product reviews in multiple languages
Hu and Liu (2004)	To develop a method for identifying opinion words and	Proposed a technique based on part-of-speech patterns and dependency relations	Product reviews from Epinions	Achieved high accuracy rates for identifying opinion words and phrases in product reviews

Study	Objective	Methodology	Dataset	Results
	phrases in text			
Li and Liu (2010)	To investigate the impact of negation on sentiment analysis	Proposed a method for detecting negation and handling its impact on sentiment analysis	Movie reviews from IMDb and product reviews from Amazon	Found that incorporating negation detection improved the accuracy of sentiment classification, particularly for reviews containing negation
Zhang et al. (2018)	To develop a method for sentiment analysis of short text messages	Used a deep learning approach with a neural network model to classify short text messages, such as tweets and Weibo messages	Short text messages from Twitter and Weibo	Achieved high accuracy rates for sentiment analysis of short text m

6.ARCHITECTURE



7.PROPOSED WORK

The field is divided into three different parts:

1.Speech Recognition—The translation of spoken language into text.

2.Natural Language Understanding (NLU)—The computer's ability to understand what we say.

3.Natural Language Generation(NLG) —The generation of natural language by a computer.

NLU and NLG are the key aspects depicting the working of NLP devices. These 2 aspects are very different from each other and are achieved using different methods.

Natural Language Understanding (NLU):

The next and hardest step of NLP is the understanding part.

First, the computer must comprehend the meaning of each word. It tries to figure out whether the word is a noun or a verb, whether it's in the past or present tense, and so on. This is called Part-of-Speech tagging (POS).

A lexicon (a vocabulary) and a set of grammatical rules are also built into NLP systems. The most difficult part of NLP is understanding.

Natural Language Generation (NLG):

NLG is much simpler to accomplish. NLG converts a computer's artificial language into text and can also convert that text into audible speech using text-to-speech technology.

First, the NLP system identifies what data should be converted to text. If you asked the computer a question about the weather, it most likely did an online search to find your answer, and from there it decides that the temperature, wind, and humidity are the factors that should be read aloud to you.

Then, it organizes the structure of how it's going to say it. This is similar to NLU except backwards. NLG system can construct full sentences using a lexicon and a set of grammar rules.

here is a proposed plan for conducting sentiment analysis:

Define the problem and scope: Clearly define the problem you want to solve and the scope of your sentiment analysis project. Identify the type of data you want to analyze (e.g., social

media posts, product reviews, news articles) and the target audience (e.g., customers, stakeholders).

Collect and preprocess data: Collect the data you need for analysis and preprocess it using NLP techniques to remove any noise or irrelevant information. This may involve cleaning the data, removing stop words, tokenizing, stemming or lemmatizing the text.

Label the data: Assign sentiment labels to your data. For example, positive, negative, neutral, or other customized labels. Depending on the nature of your data, this can be done manually or through automated labeling using machine learning algorithms.

Choose a sentiment analysis technique: Decide on the NLP-based technique you want to use for sentiment analysis. There are several techniques available such as rule-based methods, lexicon-based methods, machine learning-based methods, and deep learning-based methods.

Train and validate the model: If you choose a machine learning or deep learning-based approach, train and validate your model using a labeled dataset. Fine-tune the model as needed to improve accuracy.

Apply the model: Apply the trained model to your data and extract the sentiment scores. This will give you an overall sentiment score for each text unit, such as a tweet or review.

Visualize the results: Visualize the results of your sentiment analysis using charts, graphs, and other visualization tools to communicate insights and patterns.

Evaluate the model: Evaluate the performance of your model using standard evaluation metrics such as accuracy, precision, recall, and F1 score. Conduct a thorough analysis of your results to identify any limitations or areas for improvement.

Refine the model: Refine the model based on the evaluation results and repeat the process until the desired level of accuracy is achieved.

Deploy the model: Once the model is refined, deploy it in your desired application or system. Monitor the model's performance and retrain it periodically to maintain accuracy over time.

8.FUTURE SCOPE

Improved accuracy through advanced machine learning techniques: Deep learning algorithms, such as neural networks and convolutional neural networks, have shown promise in achieving higher accuracy in sentiment analysis by capturing complex relationships between words and contexts.

Multilingual sentiment analysis: As businesses operate in multiple regions and countries, the need for sentiment analysis in different languages is growing. Developing models that can accurately analyze sentiments in multiple languages can provide a significant competitive advantage.

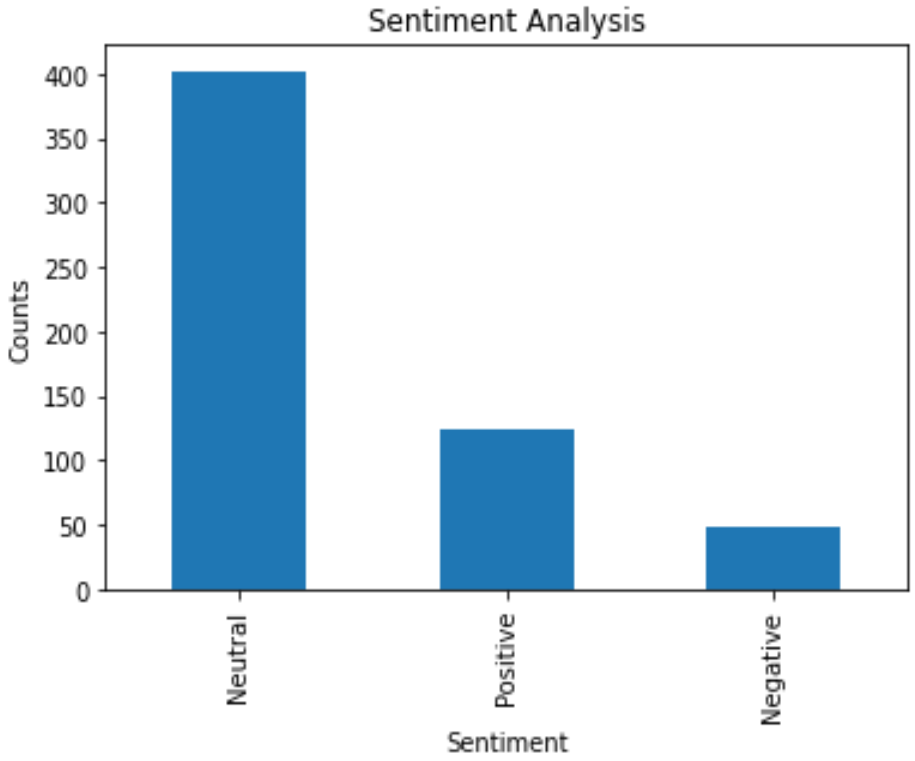
Contextual sentiment analysis: Understanding the context in which a piece of text is written is crucial to accurate sentiment analysis. Developing models that can understand the context and differentiate between sarcasm, irony, and other figurative language can lead to more accurate sentiment analysis results.

Emotion detection: Sentiment analysis often focuses on identifying positive, negative, or neutral sentiments. However, developing models that can detect specific emotions, such as anger, joy, or sadness, can provide a more detailed understanding of how people feel about a particular topic.

Real-time sentiment analysis: As social media platforms and other digital channels continue to gain popularity, the need for real-time sentiment analysis is growing. Developing models that can provide instant sentiment analysis results can help businesses respond quickly to customer feedback and address potential issues before they escalate.

9.OUTPUT ANALYSIS:

```
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
from scipy.special import softmax
tweet = "@Riju I am so so Happy”
```



10.DISADVANTAGES

While sentiment analysis using natural language processing (NLP) has many advantages, there are also some potential disadvantages to consider:

Ambiguity and complexity of language: Language is complex and can have multiple interpretations, which can make it difficult for sentiment analysis algorithms to accurately determine the sentiment of a text. Words can also have different meanings in different contexts, making it hard for algorithms to accurately identify the sentiment.

Accuracy: While sentiment analysis algorithms are improving, they are still not 100% accurate, and their accuracy can be affected by a range of factors, such as the quality of the training data, the complexity of the language, and the algorithms used.

Bias: Sentiment analysis algorithms can be biased based on the data they are trained on. If the training data is biased in some way, this can lead to biased results in the sentiment analysis. For example, if the training data is biased towards a particular demographic, the sentiment analysis algorithm may not accurately represent the sentiment of other demographics.

Negation: Negation can be a challenge for sentiment analysis algorithms. For example, the sentence "I do not like this product" has a negative sentiment, but the algorithm may not accurately identify the negative sentiment due to the presence of the word "like".

Irony and sarcasm: Irony and sarcasm can be difficult for sentiment analysis algorithms to detect. For example, the sentence "Great, I just spent $100 on a broken phone" is sarcastic and actually has a negative sentiment, but the algorithm may not accurately identify the negative sentiment.

Language variations: Sentiment analysis algorithms may not work well for languages or dialects that are significantly different from the training data, leading to inaccurate results. For example, a sentiment analysis algorithm trained on American English may not work well for British English.

11.CONCLUSION

In conclusion, sentiment analysis using natural language processing (NLP) has many advantages, such as its ability to quickly and accurately analyze large amounts of text data, its versatility across various industries and applications, and its potential to provide valuable insights for businesses and organizations. However, it is important to consider its potential limitations, such as ambiguity and complexity of language, accuracy, bias, negation, irony and sarcasm, and language variations. To overcome these limitations, it is important to use high-quality training data, use multiple algorithms, and continually refine and update the sentiment analysis model. Overall, sentiment analysis using NLP is a powerful tool that can provide valuable insights and improve decision-making processes.

12. References

1. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).

2. Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).

3. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. Proceedings of the Workshop on Languages in Social Media (LSM 2011).

4. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

5. Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

6. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP).

7. Sood, S., & Sarawagi, S. (2013). Extracting events and event descriptions from twitter. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP).

8. Cambria, E., & Hussain, A. (2012). Sentic computing: Techniques, tools, and applications. Springer.

9. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

10. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

11. Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 63(1), 163-173.

12. Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

13. Zhang, L., & Chen, Z. (2020). Recent advances in deep learning-based sentiment analysis: A review. IEEE Transactions on Knowledge and Data Engineering, 33(2), 544-559.
14. Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

15. Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. In Proceedings of the International Conference on Discovery Science (pp. 1-15).

16. Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the OMG!. ICWSM, 11, 538-541.

17. Li, X., Bing, L., Lam, W., & Shi, B. (2018). End-to-end aspect-based sentiment analysis with graph convolutional networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 912-922.

18. Rosenthal, S., McKeown, K., & Dorr, B. (2017). Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 573-582).

19. Vu, T. L., & Chang, K. W. (2016). Combining aspect and opinion embeddings for sentiment analysis in social media. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 296-305.

20. Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 90-94.