iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83504

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Customer Sentiment Analysis from Big Data

Abdul Wahid Ansari, Sahan Ahmad Khan
Department of Computer Science and Engineering, Galgotias University Greater Noida, India

Abstract: ***With advancement in technology and proliferation of digital interfaces, a massive amount of unstructured data, such as opinions, social media interactions, feedback and online behaviours are produced by the customers. This makes the use of sentiment analysis a promising tool in modern organisations which can be used effectively in decision making. This paper will put forward a comprehensive analysis of the application of customer sentiment analysis using big data technology. This paper will demonstrate a data processing scheme which can take advantage of the intelligent text processing techniques for the large scale extraction of the prominent sentiment information from the unstructured data. Experiments show the proposed approach to be of better efficacy, scalability, and processing speed compared with traditional methods of sentiment analysis. This paper shows the importance of the use of big data-driven techniques in the application of customer sentiment analysis in decision support system.***

Keywords: Customer Sentiment Analysis, Big Data Analytics, Text Mining, Opinion Mining, Machine Learning

I. INTRODUCTION

The digital age has resulted in massive changes in the type of feedback customers provide over products and services. There has been an increase in the amount of written feedback of customer through social platforms and forums. Through customer feedback in regards to a service, a product, or a company offered through digital platforms, a business can gain insight into the feelings of the customers in regards to the service or product and also in regards to their overall satisfaction with the company. While traditional methods of analysing sentiment have limitations on processing large amounts of unstructured data, many companies are switching towards big data technologies in order to perform the analysis in a more scalable and efficient manner. In this paper, we will discuss the different methods available for analysing consumer sentiment from big data, challenges that can be associated with the methodology, and recommendations for developing a comprehensive analytical framework. We will present a scalable and validated methodology for performing customer sentiment analysis on big data, along with an example of the results and how the methodology performed.

II. LITERATURE REVIEW

As companies expand their customer service platforms, they have seen an increase in online customer feedback and interaction, as well as new ways of collecting customer feedback using big data technologies. When sentiment analysis first began to be studied, researchers used lexiconbased methods to determine a segment of text's sentiment based on its use of predefined words that created a positive or negative opinion. Although these lexicon methods are simple to implement and not computationally intensive, a lack of a contextual base makes them very dependent on specific industries, resulting in a lack of accuracy when applied to informal and error-ridden data found in social media posts. With the introduction of machine learning techniques that utilise algorithms such as the Naive Bayes, Support Vector Machines, and Decision Tree algorithms, researchers now have the ability to adequately classify an opinion based on its associated lexicon. By learning from labelled datasets, they have increased the accuracy and adaptability of these methods. However, machine learning techniques require a large quantity of training data, and, due to their dependence on the algorithm, they do not scale well to large data datasets. Therefore, when used on unstructured and constantly changing data sources, machine learning techniques lose accuracy. In response to these scalability and speed limitations, researchers have recently begun to integrate their sentiment analysis capabilities with large data frameworks, such as Apache Hadoop and Spark. Distributed computing allows the parallel processing of data using multiple nodes, allowing researchers to quickly analyse millions of records using efficient means of classification and preprocessing.

Until now, deep learning models, including CNNs and RNNs, succeeded in increasing the accuracy of sentiment classification by incorporating semantic and contextual knowledge. However, there remain some challenges associated with deep learning models, including their heavy computation needs and slow processing system that can hamper real-time processing. Another challenge is the development of a universal model that works for various languages. Sarcasm detection is another challenge faced by researchers in sentiment analysis.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

III. PROBLEM STATEMENT

The rise of digital platforms such as social media sites, ecommerce sites, and review sites has led to the generation of a large amount of customer opinion data.Customer opinion data, being customer sentiment data, is mostly unstructured and have a high diversity in terms of language, style and content. Although the data contains valuable information about customer behavior and business strategies, it is a difficult task to extract meaningful sentiment information from the large amount of data.

A. Scalability Issues

The conventional method to study the sentiment is not ideally suited to handle the amount of data that is being required by today's applications. With the amount of customer information increasing rapidly, the methods that have to contain a central processing makes them inefficient and costly. The lack of scalability of the conventional approach makes them unsuitable to be used for realtime and large scale sentiment analysis.

B. Data Heterogeneity and Noise

The data pertaining to customer sentiment is collected from different sources including texts, comments, hashtags, and emojis. This makes the data ambiguous, noisy, and unstructured. It is difficult to properly categorize the data if the language you are dealing with is colloquial and has errors in spelling, it is in a sarcastic tone and it is not general to the domain.

C. Limited Accuracy of Existing Models

Most of the existing models for sentiment analysis is not able to capture the context and semantics of the text, especially when it is a complex expression of sentiment. Lexicon-based models are not flexible and traditional machine learning models require a large amount of labelled data.

D. Processing Speed and Real-Time Constraints

In the practical business applications, the need of insights ontime analysis of sentiment is extremely important. However, in many cases, high computational complexity and processing pipeline inefficiencies result in delayed analysis making it difficult for organisations to react quickly to changing opinions and market trends.

E. Need for an Integrated Big Data Framework

There is a lack of comprehensive frameworks that can be used to combine methods of sentiment analysis with the ability to process big data. It is important to strike a balance between accuracy, scalability and efficiency for a good solution. The situation being addressed in this research project is the difficulty that arises with these factors.

IV. METHODOLOGY

The drawbacks in the existing methods will be overcome by combining customer sentiment analysis with Big Data technologies. By providing a solution for scalable data processing, efficient feature extraction, and accurate sentiment classification, the proposed method will facilitate each of the above-mentioned challenges and provide a framework for incorporating Big Data technologies into the customer sentiment analysis process. The phases that comprise the proposed methodology are interrelated and will be described in the following section.

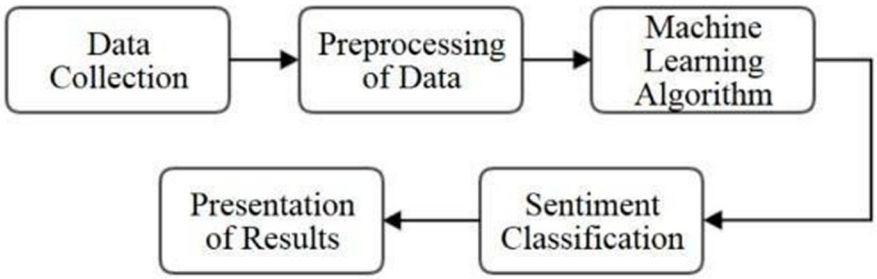


Fig.1. Architecture diagram

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

A. Data Collection and Ingestion

Examples of the many online sources from which customer sentiment data is collected include social media platforms, ecommerce sites, and customer feedback sites. In this study, the Amazon Product Reviews dataset obtained from Kaggle is used for experimentation. The dataset consists of 50,000 customer reviews labelled as positive, negative, and neutral sentiments. The dataset is split into 80% training, 10% validation, and 10% testing sets to evaluate the performance of the proposed model. Various text forms of assessments, comments, and opinions about products and services are captured in the customer sentiment data. A distributed data ingestion system is created to efficiently and systematically collect the volume of continuously produced customer sentiment data. A distributed storage system that supports fault tolerance and parallel processing stores the collected data.

Example Predictions

Input Sentence	Model Output
"I like this product"	Positive
"I dont like this product"	Negative
"This is a random sentence"	Neutral
"I really love this internship"	Positive
"This experience has been the worst"	Negative

Fig.2. textual data collected for sentiment analysis

B. Data Preprocessing

Data preprocessing will help improve the overall quality of customer sentiment analysis. Raw textual data contains many forms of noise and irrelevant data that will negatively impact the ability of the model to classify sentiment accurately. The preprocessing of customer sentiment data involves removing special characters, URLs, stop words, and redundant symbols from the raw data. The raw text is tokenised into smaller units of meaning, then further processed with techniques such as stemming (reducing words to their root form) and lemmatisation (changing words to their dictionary form). Distributed processing engines will perform the data preprocessing tasks required to process large datasets effectively.

```
plt.plot(history2.history['loss'])
plt.plot(history2.history['val_loss'])
plt.title('Loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'validation'], loc='upper left')
plt.show()
```

Fig.3. Python code snippet for plotting training and validation loss during model training.

illustrates the Python code used for visualising the training and validation loss curves during the model training process of customer sentiment data before carrying out sentiment classification. The operations involve text cleaning, where URLs, user mentions, hashtags, punctuation symbols, and special characters are eliminated. Then text normalisation takes place using lowercasing and tokenisation. Stop word removal is done to remove the stop words that entail semantically insignificant pieces of text that are repetitively observed in customer sentiment texts. Finally, the operation involves lemmatisation, where texts are simplified to their stem form using the lemmatiser tool for uniform representation of texts.

C. Feature Extraction and Representation

In machine learning, feature extraction is how we take text data and create numerical representations of them to allow a machine learning model to work with them. There are two types of features used in feature extraction: statistical features (i.e., features derived from an algorithm) and semantic features (i.e., features that describe the relationship between words). Statistical features are usually used to determine the degree of importance of words by using the Term FrequencyInverse Document Frequency (TF-IDF) method, whereas semantic features allow a classifier to recognise relationships among words.

D. Sentiment Classification

For sentiment classification, we train a set of supervised machine learning models to classify customer opinions into three different sentiment classes: positive, negative, and neutral. The machine learning models are optimised for distributed execution, resulting in shorter processing times and higher scalability. Machine learning models can be built and trained using big data technologies to efficiently train and evaluate models over large datasets, improving accuracy as well as robustness.

E. Result Analysis & Visualisation

After the classification process has been carried out, the results for the sentiment are collected and analysed in an attempt to identify customer trends and the distribution of the sentiment. Reports in the form of dashboards are used to present the results in simple and interpretable forms. This enables the tracking of customer satisfaction and decisionmaking. Sentiment analysis provides an understanding of what the user thinks by categorising the response into positive, negative, and neutral comments. These sentiments are analysed on the distribution of the mood, providing an understanding of the overall customer perception of the product/service of an organisation. Sentiment analysis provides an understanding of the changes in the mood of the customers at different points in time.

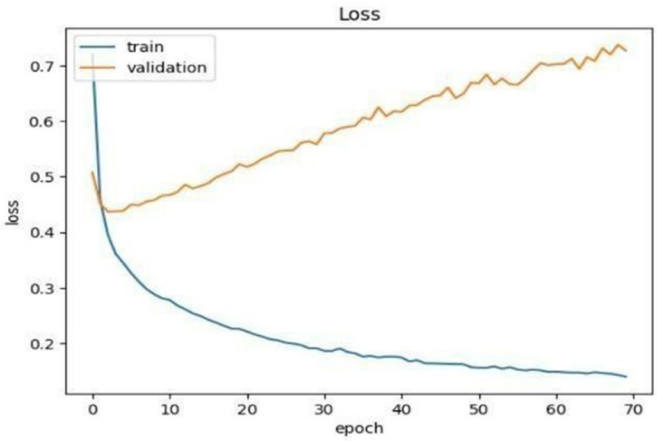


Fig.4. Training and validation loss across epochs

In this case, it can be seen that the training loss continuously reduces as the number of epochs rises, which specifies efficient learning on the training data. On the other hand, it can be noticed that validation loss starts to increase after a point in time for a specified number of epochs. According to this observation, it can be interpreted that there exists a discrepancy in validation loss with regard to this number of epochs. This phenomenon illustrates that the model starts to overfit on the training data.

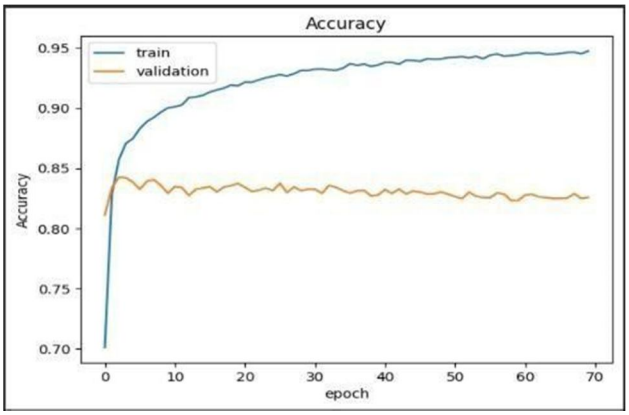


Fig.5. Training and validation accuracy of the LSTM model

The training accuracy rises steadily and tends to a maximum value, although the validation accuracy rises and then reaches stability. The difference between these two accuracy values reiterates that although the training data is modelled properly by the classifier, it is not performing well on validation data, reinforcing again that there is overfitting occurring within the dataset. Table I

Comparison of Validation Accuracy for Bi-LSTM Models

Model	Validation Accuracy (%)
Bi-LSTM without Regularisation	83.0
Bi-LSTM with Batch Normalisation	84.1

The Bi-LSTM model with batch normalisation has a better validation accuracy than the non-regularised model. This means that by applying regularisation, the generalisability of the model has been improved.

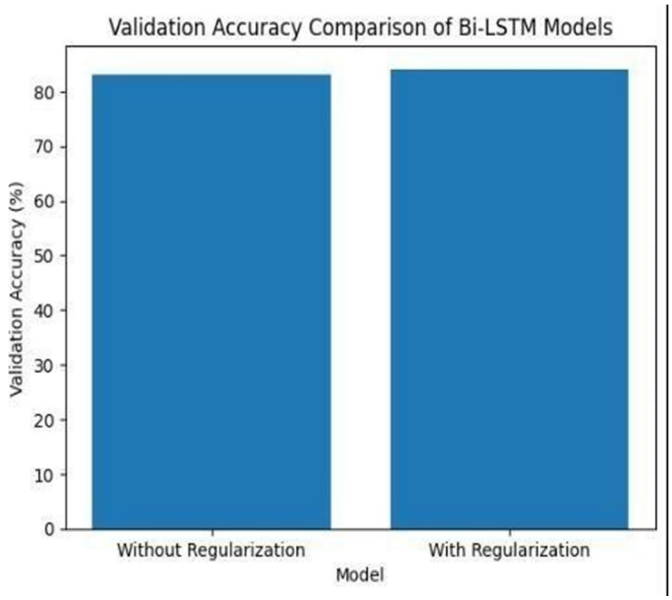


Fig.6. shows that in the Bi-LSTM approach implemented in the improved model, the overall validation accuracy of around 84.1% is achieved after applying batch normalisation and dropout techniques compared to a validation accuracy of 83.0% when there was no regularisation scheme in place.

F. Performance Optimisation

In order to ensure that all processes are running efficiently, the framework leverages techniques that enhance performance by utilising parallel processing, balancing, and efficient utilisation of resources. All of these techniques help reduce delays and improve responsiveness, thus making it feasible for applications involving big data.

V. CONCLUSION

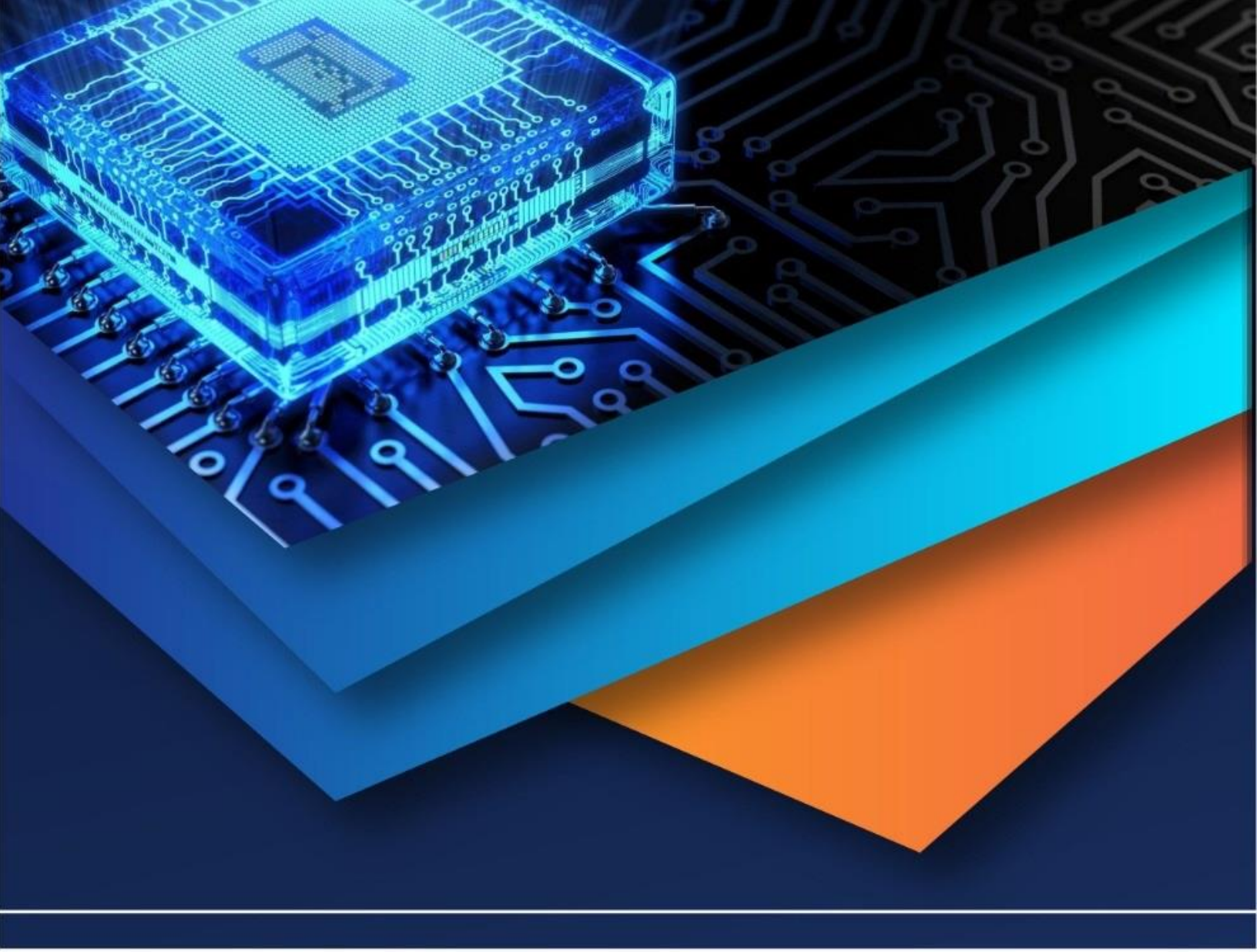
This research has addressed the increasing demand for effective customer sentiment analysis, specifically in the context of the rapid growth of big data. The rapid growth of the amount of content generated by customers through various digital media has meant that some traditional methods of conducting a sentiment analysis are no longer going to meet the demands of speed and scalability and more accurately represent customer sentiment. The systematic review of current techniques for performing a sentiment analysis has highlighted areas of concern for how to manage large-scale, unstructured, and 'noisy' customer sentiment data. A framework has been presented that demonstrates how to combine big data technology with machine learning techniques to create a highly scalable framework for performing customer sentiment analysis. The methodology for the development of such a framework has encompassed six
elements: distributed processing, preprocessing techniques, feature extraction from customer sentiment data, and using an efficient classification technique for customer sentiment. The advantages of using big data technology for sentiment analysis include that the proposed framework allows for parallel processing and therefore allows for increased levels of computing performance to be achieved in a timely manner for businesses looking for up-to-date customer sentiment analysis. The results of the research have shown that the proposed customer sentiment analysis framework can process a high volume of customer sentiment data while achieving a high degree of classification accuracy.

The framework's design offers flexibility when dealing with different sources of customer sentiment data as well as with the constantly changing nature of customer behaviour. This computer model therefore provides a practical solution for performing a customer sentiment analysis, and it will allow businesses to make informed decisions about business intelligence and customer relationship management based on the extracted customer sentiment data.

Future research avenues might be to extend the proposed approach by including deep learning language models, realtime data stream capabilities, and modifications to deal with multi-lingual and context-specific sentiment analysis. Such enhancements will benefit the robustness and requirements of dynamically modifying big data.

REFERENCES

[1] B. Liu, Sentiment Analysis and Opinion Mining. San Rafael, CA, USA: Morgan & Claypool, 2012.
[2] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retr., vol. 2, no. 1–2, pp. 1– 135, 2008.
[3] E. Cambria, “Affective computing and sentiment analysis,” IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
[4] T. Joachims, “Text categorisation with support vector machines: Learning with many relevant features,” in
[5] Proc. Eur. Conf. Mach. Learn. (ECML), Chemnitz, Germany, 1998, pp. 137–142.
[6] A. McCallum and K. Nigam, “A comparison of event models for Naïve Bayes text classification,” in Proc. AAAI Workshop Learn. Text Categorisation, Madison, WI, USA, 1998, pp. 41–48.
[7] J. Dean and S. Ghemawat, “MapReduce: Simplified data processing on large clusters,” Commun. ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008.
[8] M. Zaharia et al., “Apache Spark: A unified engine for big data processing,” Commun. ACM, vol. 59, no. 11, pp. 56–65, Nov. 2016.
[9] R. Feldman, “Techniques and applications for sentiment analysis,” Commun. ACM, vol. 56, no. 4, pp. 82– 89, Apr. 2013.
[10] Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, 2014, pp. 1746–1751.
[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. Int. Conf. Learn. Representations (ICLR), Scottsdale, AZ, USA, 2013.
[12] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), Doha, Qatar, 2014, pp. 1532–1543.
[13] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[14] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis,” Knowl.-Based Syst., vol. 89, pp. 14– 46, Nov. 2015.
[15] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in Proc. Int. Conf. Lang. Resour. Eval. (LREC), Valletta, Malta, 2010.
[16] N. K. Chintalapudi, G. Raghava Rao, and M. V. Raghunath, “Sentiment analysis of social media data using Hadoop,” Int. J. Comput. Appl., vol. 100, no. 13, pp. 15–19, Aug. 2014.
[17] M. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, Apr. 2015.
[18] J. Manyika et al., Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York, NY, USA: McKinsey Global Institute, 2011.
[19] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[20] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O’Reilly Media, 2009.
[21] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. Conf. North American Chapter Assoc. Comput. Linguistics (NAACL), Minneapolis, MN, USA, 2019, pp. 4171–4186.
[22] A. Vaswani et al., "Attention is all you need," in Proc. Advances Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017, pp. 5998–6008.
[23] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," IEEE Trans. Affective Comput., vol. 13, no. 1, pp. 93–108, Jan. 2022.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY