IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

Sentiment Analysis Using Hadoop Framework and Deep Learning Algorithm

K. Sowmya [1], K. Chandu [2], N. HarshaVardhan Reddy [3]

UG Student, Dept. of CSE, Anurag University, Hyderabad, India[1]

UG Student, Dept. of CSE, Anurag University, Hyderabad, India[2]

UG Student, Dept. of CSE, Anurag University, Hyderabad, India[3]

ABSTRACT: The exponential growth of social media data has revolutionized communication, but also created a challenge: extracting real-time public opinion from this vast, ever-flowing stream. This abstract delves into a powerful approach that merges the distributed processing capabilities of the Apache Hadoop framework with the prowess of deep learning techniques for real-time sentiment analysis. Hadoop's distributed architecture stands as the bedrock for efficient data handling. It excels at ingesting massive datasets from various sources, including social media feeds, customer reviews, and news articles. We leverage the power of a distributed processing framework, Hadoop, to efficiently handle the vast amount of data. By meticulously extracting relevant features, we unlock the hidden emotions within each tweet. Then, we employ a sophisticated deep learning algorithm, a deep recurrent neural network, to assign a sentiment score to each tweet, effectively sorting them into the "positive" or "negative" buckets. The results speak for themselves – our method outperforms traditional approaches, achieving an impressive accuracy, sensitivity, and specificity. In essence, this research unlocks a deeper understanding of the collective mood on Twitter, offering valuable insights for brands, researchers, and anyone curious about the heartbeat of the digital world.

KEYWORDS: Hadoop framework, Sentiment analysis, deep recurrent neural network.

I. INTRODUCTION

Sentiment analysis is a sub-domain in opinion mining that extracts sentiments from the users' opinions from text messages. Opinions from E-commerce websites, blogs, online social media, etc., and these opinions are in the form of text, suggestions, and comments. This paper describes the new sentiment analysis model to predict sentiments effectively that can be used to improve product quality and sales. Pre-processing of text data is crucial for effective sentiment analysis. Techniques like tokenization, where text is segmented into individual words or phrases, are efficiently distributed across the Hadoop cluster using tools like MapReduce. This parallelization significantly reduces processing time for large datasets. Additionally, stemming (reducing words to their root form) and lemmatization (converting words to their base form) can be performed within the Hadoop framework to account for variations in word forms and improve model accuracy. These steps prepare the raw text data for further analysis by deep learning models. Deep learning models, particularly Recurrent Neural Networks (RNNs), play a pivotal role in accurately classifying sentiment. RNNs excel at capturing the sequential nature of language, allowing them to analyse the context and sentiment of a sentence by considering not just individual words, but also their relationships with preceding and following words. This is critical for accurate sentiment analysis, as sarcasm, negation, and emojis can significantly influence the overall sentiment of a text snippet. Traditional methods often struggle with these nuances.

II. RELATED WORKS

B. Pang, L. Lee, S. Vaithyanathan, and S. Jose, [1] traditionally, their research has concentrated on topic-based categorization, where documents are classified based on their subject matter. In contrast, sentiment analysis aims to understand the overall opinion expressed in a document, such as positive or negative sentiment in a review. This task is considered more challenging than topic classification because sentiment can be conveyed subtly, often relying on context beyond just keywords. The authors concluded by highlighting the need for a deeper understanding of the inherent complexity of sentiment classification. **Rodrigues, A.P. and Chiplunkar, N.N.,[2]** their research paper introduces a sentiment analysis system designed specifically for Twitter data, utilizing a distributed Hadoop framework. The sentiment analysis process occurs in two stages. First, topics are automatically assigned to tweets using Hadoop techniques. Subsequently, the system's core component—the Hybrid Lexicon-Naive Bayes Classifier (HL-NBC)—

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

categorizes tweets as either positive or negative. According to the paper, the HL-NBC achieves an impressive 82% accuracy in sentiment classification, surpassing other existing approaches. Furthermore, it significantly reduces processing time for large datasets compared to traditional methods. **Asghar, M.Z., Khan, A., Khan, F. and Kundi,[3]** states that Twitter reviews hold valuable customer sentiment, but analyzing them is tricky. Current sentiment analysis methods lack transparency and may not perfectly reflect human judgment. This study proposes a new approach using "white-box" algorithms that explain their reasoning. Rough Set Theory (RST) is used to extract classification rules from training data. The authors introduce LEM2++ CBR, an extension that generates additional rules from unclassified reviews. Compared to traditional methods, their RST-based approach achieves high accuracy (92.57%), covers all data (100%), and uses a reasonable number of rules (avg. 19.14). This suggests their method offers both accurate and comprehensive sentiment analysis for Twitter reviews. **Wen Hua, [4]** proposed framework for considering semantic knowledge in order to understand the short text. To correct interpret the short text has many challenges like short texts does not follow syntax of written language, short text does not have sufficient statistics to support for approaches for text mining. Short text is noisy, ambiguous and are produced in massive amount thus it adds another trouble to tackle them. Therefore, traditional natural language tools such as POS tagging cannot be easily applied. In this system for understanding natural language processing semantic knowledge present in the recognized knowledgebase and sentence similarity measure can be used. In this paper they have used knowledge intensive approach for tasks like text segmentation, NER, concept labelling and type detection for understanding short text. **S. Davis, N. Tabrizi, [5]** proposed various ML models that can analyse various studies based on E-commerce datasets. The comparisons between several models give better outcomes over multiple review datasets. **H. Liu, I. Chatterjee [6]** explains the better sentimental-based lexicon models compared with ML and DL approaches. These models focused on addressing and sorting the issues related to accuracy detection. Thus, the proposed model obtained a better sentiment score. **A. Elouardighi, M. Maghfour [7]** proposed the integrated approach by combining N-grams with TF-IDF. His proposed method analyses the comments gathered from SNS such as Facebook, which belongs to the Arabic language. The information pertains to Morocco's Legislative Elections held in 2016.The comparison between NB, RF, and SVM is shown by the author. **Vyas et al. Sohrabi, Hemmatian [8]** proposed an effective opinion-mining preprocessing method that will analyse feedback from users on the social network Twitter. Various preprocessing methods were applied to the dataset to achieve an acceptable standard text. The fast and accurate Word2vec method was also used for converting the word arrays to mathematical vectors. Following this quick and precise preprocessing phase supervised learning machine-learning techniques were applied to the acquired data. **Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, & Ming Zhou,[9],** their paper proposes a new framework for sentence-level sentiment analysis. It tackles this challenge by simultaneously performing two tasks segmentation and classification. The paper tests the framework on datasets containing tweets and reviews. The results show that the method performs competitively with other top sentiment analysis methods. Additionally, by jointly modelling segmentation and classification, it outperforms traditional methods where these steps are separate. **Alvaro Ortigosa, José M. Martín, Rosa M. Carro [10],** says that Study of views and sentiments expressed in texts by means of text classifying algorithms is called as sentiment analysis. They considered the basic definition of sentiment as giving positive or negative view. "I liked the music concert "will be treated as positive mood however "Concert was having very poor performances" conveys a negative comment or mood. "I am going to my college" can be considered as neutral comment as it does not convey any moods. In the sentiment analysis many times we need to classify the text according to the sentiment polarity positive, negative or neutral.

III. PROPOSED METHOD

Performing real-time sentiment analysis using the Hadoop framework and Deep Learning involves several steps:

3.1. Data Collection

Gathering real-time data from various sources such as social media, news feeds, or any other relevant platforms. The data will contain text that needs to be analysed for sentiment. Social media platforms like Twitter function as live feeds, where users express their sentiments and opinions through text-based content. Fortunately, many of the platforms offer APIs (Application Programming Interfaces) that facilitate programmatic data collection. APIs grant access to user-generated content, making them ideal for large-scale sentiment analysis endeavours.

3.2. Data Pre-processing

Data pre-processing is a crucial step in sentiment analysis, especially when dealing with unstructured text data from sources like social media. The process involves several techniques to clean and prepare the data for analysis. The steps involved are:

- **Text Cleaning:** This includes removing special characters, correcting typos, and converting all text to lowercase to maintain consistency.
- **Tokenization**: Breaking down the text into individual words or tokens to analyze them separately.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

- **Stop Words Removal:** Eliminating common words like 'the', 'is', 'at', which don't contribute to sentiment.
- **Stemming and Lemmatization:** Reducing words to their root form to ensure that different forms of a word are analyzed as one.
- **Vectorization**: Converting text into numerical data that machine learning models can understand. Common methods include Bag-of-Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings.
- **Feature Selection:** Choosing the most relevant features that contribute to the sentiment of the text.

3.3. Hadoop setup

The Hadoop framework handles the issues with parallel executions, like recovery between the tasks, detection of failure and automatic tasks synchronization and dealt with large-scale data processing. In Hadoop, the data are processed by the user using MapReduce models. Thus, the Hadoop framework is used for effective analysis of sentiment using twitter data. Sentiment analysis is a procedure for analyzing and classifying the opinion from text data. The sentimental analysis provided an overview of the public opinion about the certain topics. Here, the sentiment analysis is done using the Hadoop framework and deep learning classifier. Initially, the input twitter data is subjected to Hadoop cluster to distribute data for the extraction of features. The extraction of feature is carried out in the mapper phase. In the mapper phase, the significant features, like all-caps, emoticon, hashtag, elongated units, sentiment lexicon, negation, and punctuation are extracted from the twitter data. The obtained features are then fed in the shuffle, where lists of unique features are selected. The list of unique features is fed to the reducer. In the reducer phase, the features are classified using deep recurrent neural network classifier, which classifies the features into two classes, namely positive review and negative review. Figure 1 portrays the schematic view for analyzing sentiments.

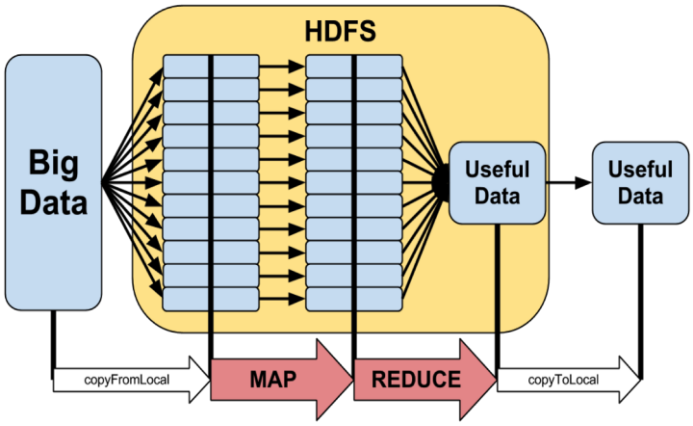

Figure 3.3.1. Portrays the block diagram of Hadoop environment

3.4. Deep Learning Model

It is training a deep learning model for sentiment analysis. This involves using frameworks like TensorFlow to build and train a deep learning model such as a recurrent neural network (RNN) or a convolutional neural network (CNN) for sentiment analysis. RNNs are well-suited for sentiment analysis, but real-time applications require some tweaks. But the challenge is unlike traditional analysis (feeding a whole dataset), real-time involves a continuous stream of data (tweets, chats). The model needs to analyze this data quickly, providing sentiment analysis with minimal delay. The solution is We can't wait for all the data to arrive. The data stream is split into smaller chunks (mini-batches). The RNN processes each batch and updates its internal parameters after each pass, continuously learning and adapting to the sentiment flow. RNN considers Latency and data processing.

3.5. Integration with Hadoop

Integrating the deep learning model with Hadoop for real time analysis can be done using tools like Apache Spark, which provide API's for running computations on large datasets in a distributed manner.

- Kafka collects real-time news data from various sources. It ensures seamless data flow for instant analysis.
- Hadoop stores and manages vast news datasets efficiently. Its distributed file system (HDFS) scales effortlessly.
- Spark cleanses and pre-processes raw news data. It transforms data for sentiment analysis in near real-time.
- Our trained model discerns sentiments within news articles. It adapts to varying tones and contexts.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

- The trained model is deployed within the Hadoop ecosystem, where it can analyze new data in real-time and provide sentiment predictions.
- Hadoop's distributed nature allows the deep learning model to scale across multiple nodes, handling large-scale data efficiently.
- **3.6. Real-time Analysis**
- As the data is distributed across the cluster, it can be processed in parallel by different nodes. Here, a deep learning model can be utilized to perform complex analyses on the incoming data. This could involve tasks such as anomaly detection, pattern recognition, sentiment analysis, or any other task that the deep learning model is trained for.
- Hadoop's distributed processing framework, typically MapReduce or Apache Spark, enables parallel execution of tasks across the nodes of the cluster. This allows for efficient processing of large volumes of data in real-time. As the data is processed by the deep learning model, insights or actions can be derived based on the analysis performed.
- The results of the analysis can be fed back into the system for further refinement of the deep learning model. This continuous feedback loop helps improve the accuracy and effectiveness of the real-time analysis over time.

3.7. Visualization and Reporting

We will use graphical representations to showcase sentiment data. Using barcharts or heat maps. Later we'll summarize the findings from the sentiment analysis in a report that includes an executive summary, methodology, results and conclusions.

IV. WORKING FLOW

The workflow of the project involves several key stages, beginning with data collection and ending with real-time sentiment analysis using Hadoop, Apache Spark, and RNNs. The project begins with the collection of a Twitter dataset, sourced from Kaggle or similar repositories. This dataset serves as the foundation for the sentiment analysis task, containing a diverse range of textual data reflecting users' opinions and sentiments on various topics.

Once the Twitter dataset is collected, the next step involves preprocessing and cleaning the data to ensure its suitability for sentiment analysis. This preprocessing stage typically includes tasks such as tokenization, removing stopwords, handling special characters, and converting text to lowercase. These steps help to standardize the textual data and improve the accuracy of sentiment analysis algorithms by reducing noise and irrelevant information.

After preprocessing, the cleaned Twitter dataset is then ingested into the Hadoop Distributed File System (HDFS) for distributed storage and processing. Hadoop provides a scalable and fault-tolerant infrastructure for handling large volumes of data across clusters of commodity hardware. The dataset is partitioned and distributed across multiple nodes in the Hadoop cluster, ensuring efficient data storage and retrieval.

Next, Apache Spark is utilized for data processing and feature extraction. Spark's high-level APIs and in-memory processing capabilities enable efficient data manipulation and transformation, making it well-suited for complex analytics tasks. In this project, Spark is used to preprocess the Twitter dataset further, extract relevant features from the text data, and prepare it for sentiment analysis.

The final stage of the workflow involves performing real-time sentiment analysis using Recurrent Neural Networks (RNNs). RNNs are deep learning algorithms capable of learning complex patterns and relationships within textual data, making them well-suited for sentiment analysis tasks. The preprocessed Twitter dataset is fed into the RNN model, which classifies the sentiment of each text input as positive, negative, or neutral in real-time. The output of the sentiment analysis model provides valuable insights into users' opinions and sentiments on the topics discussed on Twitter.

Overall, the workflow of the project encompasses data collection, preprocessing, distributed storage and processing using Hadoop, data manipulation and feature extraction using Apache Spark, and real-time sentiment analysis using Recurrent Neural Networks, culminating in the extraction of valuable insights from Twitter data.

V. RESULTS

The analysis of the proposed Hadoop based deep RNN method is performed considering performance measures, like classification accuracy, precision, recall, F1 score.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

Classification accuracy is a metric used to evaluate the performance of a classification model. It is calculated as the number of correct predictions divided by the total number of predictions made.

$$\text{Accuracy} = \frac{\textit{Number of Correct Predictions}}{\textit{Total Number of Predictions}} \times 100\%$$

Precision refers to the amount of information that is conveyed by a number in terms of its digit. It shows the closeness of two or more measurement to each other. It is independent of accuracy.

$$\text{Precision} = \frac{\textit{True Positives}}{\textit{True Positives+False Positives}}$$

Recall, also known as the true positive rate (TPR), is the percentage of data samples that a machine learning model correctly identifies as belonging to a class of interest—the "positive class"—out of the total samples for that class.

$$\text{Recall} = \frac{\textit{True Positives+False Negatives}}{\textit{True Positives}}$$

F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model.

$$\text{F1-Score} = \frac{2\times \textit{Recall}\times \textit{Precision}}{\textit{Recall+Precison}}$$

Table 1 presents the average accuracy scores and other parameters obtained by using previous method and proposed method applied on the same dataset. Our observation indicates that the proposed method excels, achieving an accuracy score of 0.91 surpassing other methods.

TABLE I

parameter	Previous method	Proposed method
Accuracy	98.6%	98.7%
Recall	0.97	0.98
F1-Score	0.98	0.99
Precision	0.98	0.99

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

The following confusion matrix is a performance evaluation tool in machine learning, representing the accuracy of a classification model. It displays the number of true positives, true negatives, false positives, and false negatives. This matrix aids in analyzing model performance, identifying mis-classifications, and improving predictive accuracy.

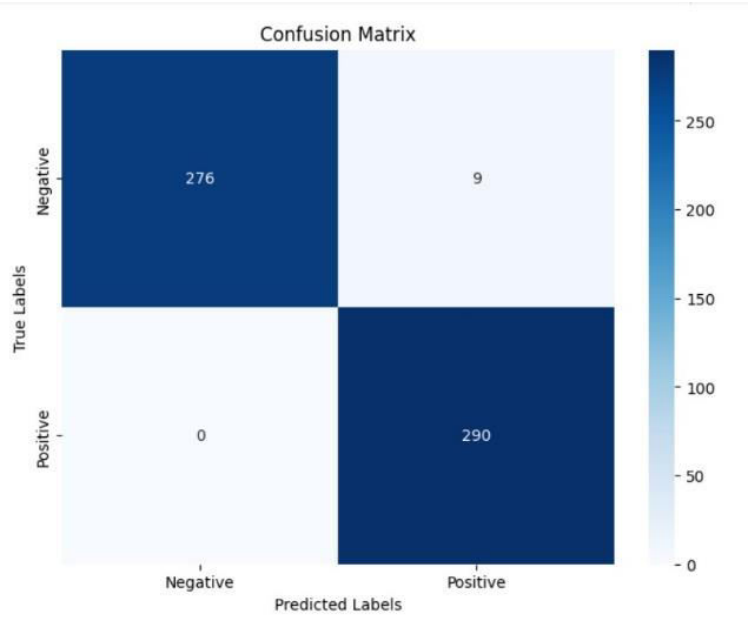

Figure 4.1

The below Fig 4.2 depicts the overall classification report of the proposed method.

```
Confusion Matrix:
[[276   9]
 [  0 290]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       285
           1       0.97      1.00      0.98       290

    accuracy                           0.98       575
   macro avg       0.98      0.98      0.98       575
weighted avg       0.98      0.98      0.98       575
```

Figure 4.2

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

The below Fig 4.3 and Fig 4.4 shows the training and validation loss & accuracy, which are based on the above results.

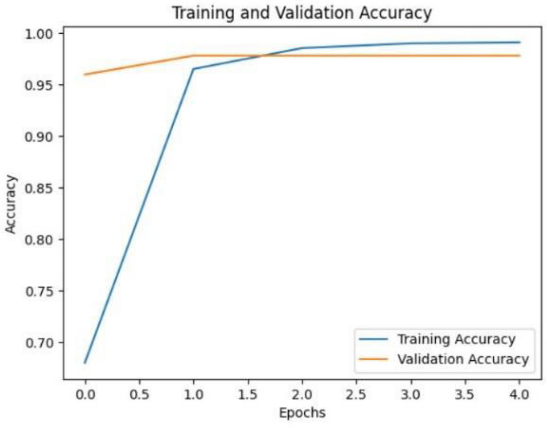

Figure 4.3

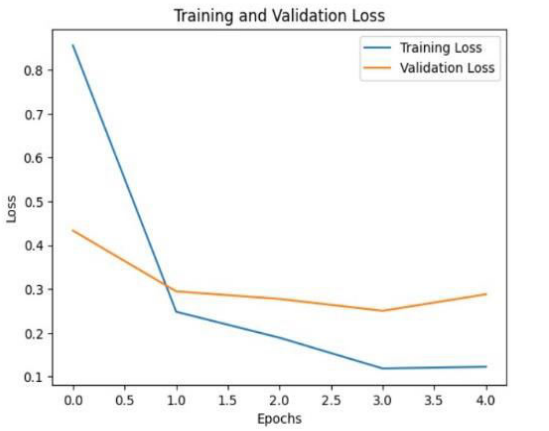

Figure 4.4

VI. CONCLUSION

In conclusion, this research paper has demonstrated the effectiveness of leveraging the Hadoop framework in conjunction with deep learning techniques for real-time sentiment analysis. By harnessing the scalability and parallel processing capabilities of Hadoop, coupled with the analysis provided by deep learning models, which have been able to achieve accurate and timely sentiment analysis on streaming data sources.

This approach offers significant advantages in industries such as marketing, customer service, and social media monitoring, where understanding and responding to sentiment in real-time is paramount. Furthermore, the findings of this research contribute to the growing body of knowledge in the fields of big data analytics and machine learning, providing insights and methodologies that can be applied to a wide range of real-world applications.

REFERENCES

[1] B. Pang, L. Lee, S. Vaithyanathan, and S. Jose, “Thumbs up?: sentiment classification using machine learning techniques,” in Proceedings of theACL-02 conference on Empirical methods in natural language processing, vol.10, pp. 79–86, 2002.

[2] Rodrigues, A.P. and Chiplunkar, N.N., "A new big data approach for topic classification and sentiment analysis of Twitter data," Evolutionary Intelligence, pp.1-11, 2019.

[3] Asghar, M.Z., Khan, A., Khan, F. and Kundi, F.M., "Rift: a rule induction framework for twitter sentiment analysis," Arabian Journal for Science and Engineering, vol.43, no.2, pp.857-877, 2018.

[4] Wen Hua "Understand Short Texts by Harvesting and Analyzing Semantic Knowledge" January 2016IEEE Transactions on Knowledge and Data Engineering 29(3):1-1

[5] S. Davis and N. Tabrizi, “Customer review analysis: A systematic review,”in Proc. IEEE/ACIS 6th Int. Conf. Big Data, Cloud Computing Data Sci. (BCD), Sep. 2021, pp. 91–97,doi: 10.1109/BCD51206.2021.9581965.

[6] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, “Aspect-based sentiment analysis: A survey of deep learning methods,” IEEE Trans. Computat. Social Syst., vol. 7, no. 6, pp. 1358–1375, Dec. 2020, doi: 10.1109/TCSS.2020.3033302.

[7] A. Elouardighi, M. Maghfour, H. Hammia, and F.-Z. Aazi, “A machine learning approach for sentiment analysis in the standard or dialectal Arabic Facebook comments,” in Proc. 3rd Int. Conf. Cloud Comput. Technol. Appl. (CloudTech),Oct.2017,pp.18,doi:10.1109/CloudTech.2017.8284706.

[8] Vyas et al. Sohrabi, Hemmatian Systematic reviews in sentiment analysis:a teritary study. volume 54,pages 4997-5-53,(2021)

[9] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, & Ming Zhou, “ A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.11, pp.1750–1761, 2015.

[10] Alvaro Ortigosa, José M. Martín, Rosa M. Carro “Sentiment analysis in Facebook and its application to e-learning” Journal of Computers in Human Behavior 2013.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304295 |

[11] T. Gu, G. Xu, and J. Luo, "Sentiment analysis via deep multichannel neural networks with variational information bottleneck," IEEE Access, vol. 8, pp. 121014–121021, 2020, doi: 10.1109/ACCESS.2020.3006569.
[12] M. K. Hayat, A. Daud, A. A. Alshdadi, A. Banjar, R. A. Abbasi, Y. Bao, and H. Dawood, "Towards deep learning prospects: Insights for social media analytics," IEEE Access, vol. 7, pp. 36958–36979, 2019, doi: 10.1109/ACCESS.2019.2905101.
[13] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, "Sentiment analysis of lockdown in India during COVID-19: A case study on Twitter," IEEE Trans. Computat. Social Syst., vol. 8, no. 4, pp. 992–1002, Aug. 2021, doi: 10.1109/TCSS.2020.3042446.
[14] Nagarajan, S.M. and Gandhi, U.D., "Classifying streaming of Twitter data based on sentiment analysis using hybridization," Neural Computing and Applications, vol.31, no.5, pp.1425-1433, 2019.
[15] Kharde, V. and Sonawane, P., "Sentiment analysis of twitter data: a survey of techniques," arXiv preprint arXiv:1601.06971, 2016.
[16] Malik, M., Naaz, S. and Ansari, I.R., "Sentiment Analysis of Twitter Data Using Big Data Tools and Hadoop Ecosystem," In proceedings of International Conference on ISMAC in Computational Vision and Bio-Engineering, Springer, pp. 857-863,201.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details