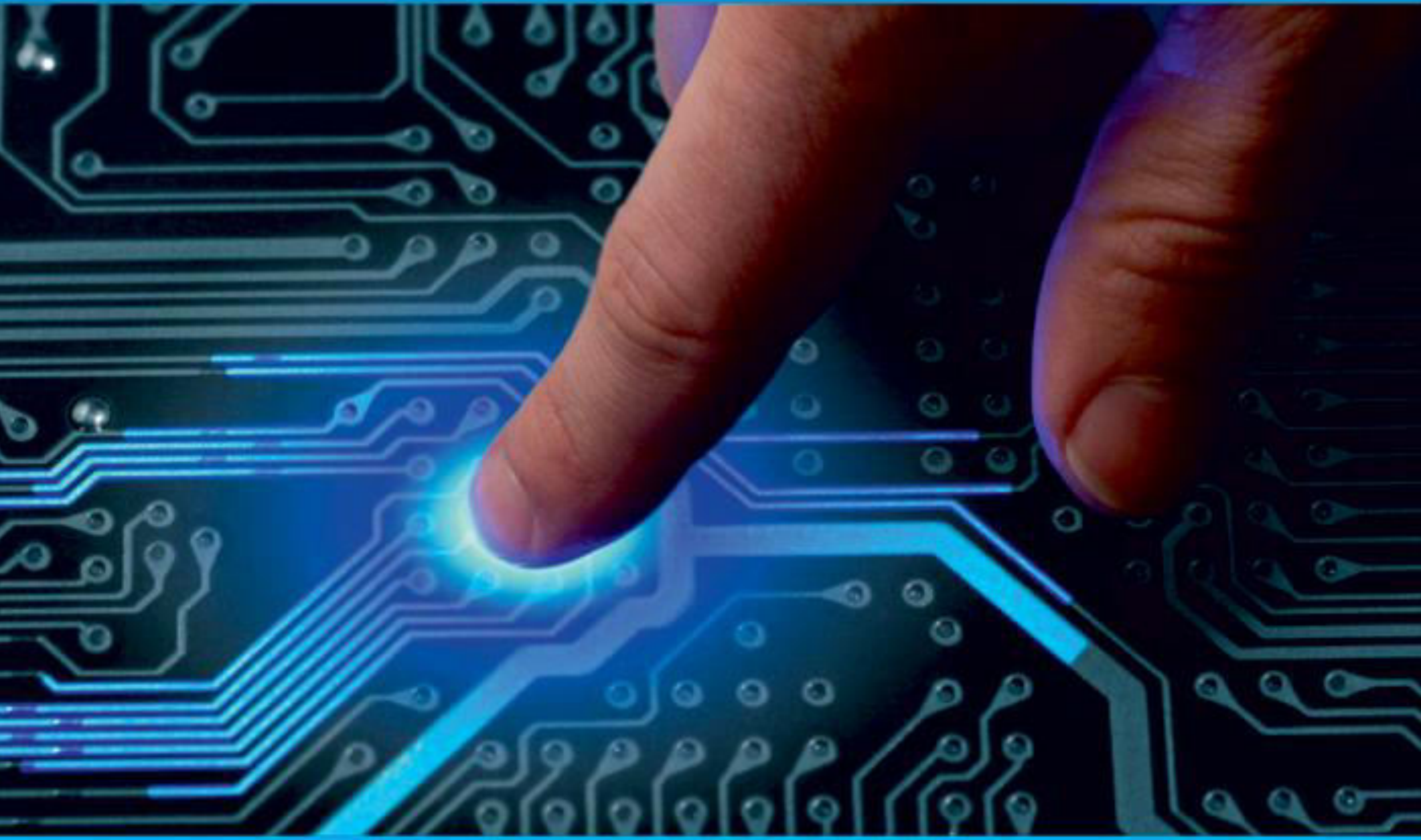
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Analysis of Different Methods for Sentiment Analysis

Onkar Rajesh Kolvankar[1], Prof. Swati Chopade[2], Prof. Sandeep Chopade[3]

MCA Student, Department of MCA, VJTI Matunga, Mumbai, India [1]

HOD, Department of MCA, VJTI Matunga, Mumbai, India [2]

Assistant Professor, Department of Mechanical Engineering Department, KJSCE, Mumbai, India [3]

ABSTRACT: This paper aims to investigate the efficiency of various machine learning algorithms in sentiment analysis based on a part of the IMDB movie reviews data set. We compare five traditional models: The four selected algorithms are Naive Bayes, Logistic Regression, Support Vector Machines (SVM), Decision Tree and Random Forests. For each model, accuracy, confusion matrix, and classification report are used as the criteria for the model assessment. The dataset includes 5000 movie reviews randomly selected from the online source which are cleaned to remove noise and other unwanted data. The textual input is converted to numerical features by using CountVectorizer. Naive Bayes model runs fast and effectively because of its simplicity, while Logistic Regression provides accurate and interpretable results. The model that performs the best is the Logistic Regression with the highest accuracy, which proves the model's efficiency in high-dimensional data. This type of comparison helps to identify the advantages and disadvantages of each model and make decisions about which models to use for further research in sentiment analysis depending on the trade-offs between complexity and accuracy, interpretability and effectiveness.

I. INTRODUCTION

Opinion mining, also known as sentiment analysis, is an important branch of natural language processing, which is aimed at the identification and processing of subjective data from the text. It is critical in gauging the perception of the public, customers, and the general trends in the society for various purposes like marketing, product design, financial analysis, and social media analysis. Due to the rapid increase in the volume of data on the web and user-generated content, the use of sentiment analysis has become critical for organizations, governments, and scholars who need to extract information from large textual data.

The purpose of this research paper is to review various techniques used for sentiment analysis and apply them on the IMDB movie reviews dataset. Using Naive Bayes, Logistic Regression, SVM, Decision Trees, and Random Forests, we aim to determine how well these techniques work in classifying movie reviews as positive or negative based on the sentiments found in the text.

The IMDB dataset is one of the most popular datasets used in the sentiment analysis domain, containing a vast number of movie reviews with their corresponding positive/negative sentiments. Every review is informative in terms of the viewers' attitude and perception towards a specific movie, which makes it suitable for sentiment analysis. The first research question of this study is to compare the efficiency of various machine learning algorithms in sentiment analysis and to reveal the advantages and limitations of each method. The experiments comparing Naive Bayes, Logistic Regression, SVM, Decision Trees, and Random Forests based on accuracy, robustness, and computational efficiency will help to understand which of the methods is more suitable for sentiment analysis.

Motivation:
The rationale for this research is based on the fact that the need for efficient sentiment analysis solutions is growing rapidly across different sectors and uses. Sentiment analysis helps the business to know what the customers are saying about their products, the general attitude of the market towards their products and make decisions on the future product development, marketing strategies and brand management. Likewise, sentiment analysis helps policymakers and researchers to understand the people's opinion, track social trends, and evaluate the effects of policies and events on the overall mood of society.

Although there are a plethora of machine learning algorithms and techniques that can be utilized for sentiment analysis, it is imperative to assess and compare these methods for practical applications. Through the empirical evaluation of Naive Bayes, Logistic Regression, SVM, Decision Trees, and Random Forests on the IMDB dataset, the study intends to offer practical recommendations and findings to practitioners and researchers regarding the effectiveness and usability of the examined techniques for sentiment analysis tasks.

Research Objectives:

1. The objectives of this study are as follows: To assess and compare the effectiveness of Naive Bayes, Logistic Regression, Support Vector Machine, Decision Trees, and Random Forests in sentiment analysis using the IMDB movie reviews dataset.
2. To measure the efficiency of each method in classifying movie reviews as positive or negative sentiments, and to determine the accuracy, precision, recall, and F1-score.
3. To compare the performance of each machine learning algorithm in sentiment analysis and determine the advantages and disadvantages of each method and guidelines on how to choose the right one depending on the problem at hand and available resources.
4. For the purpose of comparing the computational complexity of each of the methods, which include training time, memory requirements, and size of the models.

II. LITERATURE SURVEY

A Review on Sentiment Analysis Methodologies, Practices, and Applications

Sentiment analysis is a branch of natural language processing (NLP) that has received much attention in the recent past because of its versatility across many disciplines including marketing, customer feedback, social media analysis, and recommender systems. This review presents a systematic categorization of the existing approaches, methods, and uses of sentiment analysis.

The review categorizes sentiment analysis approaches into three main categories: The following are the general categories of methods: lexicon-based methods, machine learning-based methods, and the combination of both. The lexicon-based approaches use sentiment lexicon or dictionary that contains words with polarity scores assigned to them. These methods are easy to understand and implement but may not cover all the information and can depend on the context. While, the machine learning methods that learn the sentiment patterns from the labelled data are Naive Bayes, Logistic Regression, Support Vector Machines, and deep learning models such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs). Hybrid methods are more effective than the other two approaches since they incorporate the best features of both the lexicon-based and machine learning-based approaches.

The review discusses various practices and challenges associated with sentiment analysis, including data pre-processing, feature selection, model evaluation, and domain adaptation. It highlights the importance of domain-specific sentiment lexicons, feature engineering techniques, and ensemble learning methods for improving sentiment analysis accuracy and robustness.

As for the application, the most common one is the social media monitoring, in which sentiment analysis allows companies to monitor the public opinion on their brands, products, and services in real-time. It also helps in the development of sentiment-sensitive marketing initiatives, advertisement placement, and reputation management. It is also useful in the financial markets for the purpose of forecasting movement of the stock market by sentiment in the articles and tweets.

In conclusion, this review is informative and helpful to identify the current trends, approaches, and case studies in sentiment analysis and to guide the future directions of research and innovation in this area.

A Survey on Sentiment Analysis Methods, Applications, and Challenges

This survey aims to give a detailed description of the research on sentiment analysis techniques, uses, and limitations, as well as the new developments in this area. The survey is comprehensive and discusses topics such as the sentiment analysis techniques, sentiment lexicons, datasets resources, evaluation metrics and application areas.

The survey categorizes sentiment analysis methods into three main approaches: The first class of approaches is the lexicon-based methods, the second class is the machine learning-based methods, and the third class is the hybrid methods. Lexicon-based methods utilize sentiment lexicons or dictionaries that include words accompanied by polarity scores. These methods are easy to implement and explain but they do not necessarily account for context. On the other hand, Machine learning based approach learns the sentiment patterns from the labeled data and includes the algorithms like Naive Bayes, Logistic Regression, Support Vector Machines and deep learning models like Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) Networks. Hybrid methods are developed as an attempt to use the advantages of both lexicon-based and machine learning-based methods to enhance the performance of sentiment analysis.

The survey also highlights various uses of sentiment analysis in different fields such as social media analysis, customer feedback analysis, market analysis, and political analysis. It emphasizes the role of sentiment analysis in gathering information about the population's opinion, detecting trends, and making decisions based on the collected data.

In terms of challenges, the survey identifies issues such as data sparsity, domain adaptation, sentiment ambiguity, and sarcasm detection. It also discusses the limitations of existing sentiment analysis techniques and the need for more robust and context-aware models.

Overall, this survey provides a comprehensive overview of sentiment analysis methods, applications, and challenges, serving as a valuable resource for researchers, practitioners, and industry professionals in the field of natural language processing and computational linguistics.

Machine Learning Techniques for Sentiment Analysis: A Review

Turning to the specificity of the topic, this review discusses different facets of the machine learning for sentiment analysis and offers a comprehensive evaluation of various approaches, algorithms, and methodologies in the field. Based on the selected approach, the work presents both classical machine learning algorithms and models based on deep learning architectures, their advantages and disadvantages, as well as a number of tasks that can be solved using sentiment analysis.

The review first explains popular machine learning classifiers like Naive Bayes, Logistic Regression, Support Vector Machines, decision trees and consequently goes on to explain DL methods like NNs and CNNs. Sentiment lexicon based algorithms have become popular especially because they are easy to implement, not time-consuming and their results can be easily explained. The review also defines some of the feature engineering techniques such as the bag-of-words, n-gram formation, and TF-IDF which are very important in the preparation of the text data for feeding into the machine learning models.

Since sentiment analysis deals with the characters' sentiments, the deep learning models have been receiving impressive attention over the years given their potential for learning even complex patterns directly from the raw data. This paper describes the basic and more advanced architectures applied in the sentiment analysis using deep learning techniques, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and attention mechanisms. It also looks at recent advances like BERT and GPT that use pretraining language models that are currently performing near optimal on sentiment analysis tasks through using large texts and transfer learning. Some of the strengths and weaknesses of the machine learning techniques applied to sentiment analysis were identified, such as the need for the quality of data, different features of the representation, selection of the model, and finally the metrics to evaluate the model. It also explains novel R&D themes and future developments in the topic, like multi-modal sentiment analysis, aspect-oriented sentiment analysis, and sentiment analysis for low-semantic-resource languages.

In conclusion, this review gives a broad outline of the current research in the sentiment analysis of texts using machine learning techniques, which will be useful for researchers and practitioners to comprehend the current cutting-edge methods used in the analyzing of texts and the application of the best practices in this field.

A Survey on Feature Level Sentiment Analysis

This survey targets feature-level sentiment analysis that seeks to capture sentiment-carrying feature or aspect of a text. Feature-level sentiment analysis is needed when more specific level of sentiment analysis is required, which is the case in aspect-based sentiment analysis and opinion mining.

The survey categorizes feature-level sentiment analysis techniques into three main categories: It is classified mainly into lexicon-based approaches, machine learning approaches, and the combination of both the approaches. It is based on sentiment lexicons or dictionaries which contain words and phrases with polarity scores. These methods select sentiment-bearing features by using the lists of predefined keywords or phrases in the text. In contrast, machine learning-based approaches train models to recognize sentiment-carrying features from labeled data by employing supervised learning frameworks like Naive Bayes, Logistic Regression, and Support Vector Machines. The idea of using hybrid is to integrate the advantages of both lexicon-based and machine learning-based methods to enhance the feature level sentiment analysis.

Some of the challenges and future works covered in the survey are, Aspect extraction, sentiment classification, context modelling and opinion summarization. It also emphasizes the significance of domain knowledge and the techniques of domain adaptation for enhancing the accuracy and reliability of feature level sentiment analysis. In summary, this survey offers a state-of-the-art review of feature-level sentiment analysis methodologies, their use cases, and limitations, which can be a great reference for researchers, developers, and other stakeholders engaged in fine-grained sentiment analysis and opinion mining tasks.

Text-Based Sentiment Analysis Using LSTM
This paper focuses on text-based sentiment analysis using long short-term memory (LSTM) networks, a type of recurrent neural network (RNN) capable of capturing long-range dependencies in sequential data. LSTM networks have gained popularity in sentiment analysis tasks due to their ability to effectively model text sequences and capture semantic information.

The paper begins by providing an overview of LSTM networks and their architecture, highlighting their key components such as input gates, forget gates, and output gates, which enable them to learn and retain information over long sequences. It then discusses various approaches for text-based sentiment analysis using LSTM networks, including word-level and character-level representations, attention mechanisms, and transfer learning techniques.

The paper explores different datasets and evaluation metrics commonly used in text-based sentiment analysis tasks, including accuracy, precision, recall, and F1-score. It also discusses practical considerations such as data pre-processing, model training, and hyperparameter tuning for achieving optimal performance.

In terms of applications, the paper highlights the use of text-based sentiment analysis in various domains, including social media monitoring, customer feedback analysis, and product recommendation systems. It discusses real-world case studies and applications where LSTM networks have been successfully applied to extract sentiment from textual data and make data-driven decisions.

Overall, this paper provides a comprehensive overview of text-based sentiment analysis using LSTM networks, offering insights into the state-of-the-art techniques, best practices, and practical considerations in the field. It serves as a valuable resource for researchers, practitioners, and industry professionals interested in leveraging deep learning techniques for sentiment analysis tasks.

III. PROPOSEDMETHODOLOGY/PROJECT IMPLEMENTATION

The approach towards achieving the proposed methodology and the execution of the sentiment analysis project on the IMDB movie reviews dataset involves various steps collection of data, pre-processing the data, transforming the data, selecting the appropriate model, training the model, and assessing the results. Undefined

Data Preprocessing:
Data pre-processing comes first in the process of carrying out any project and in this particular case, it entails washing of the text data. This also involves the process of converting the text to lowercase, where all the punctuations are removed, all the HTML tags are stripped off from the text and all the other non- alphanumerical values are completely eliminated from the text as well as completely removing all the stopwords. Further, the dataset is bemused so that a small amount can be taken for quick tests and model checking for efficient computational operations.

Feature Engineering:
After the data has been pre-processed, feature engineering method are used to transform the textual data into features that may be used in a machine learning algorithm. Here, in this project, we utilize the CountVectorizer class from the scikit-learn library which helps in converting the text data into token counts. Since the vectorization is performed, the frequencies of words in the given documents are also preserved so that the final machine learning models can learn from the patterns and make predictions.

Model Selection:
Once the features have been extracted, next step is Model selection which involves testing several machine learning algorithms for their suitability in sentiment analysis. In this project, we focus on three popular models: Including dtree algorithms such as Naive Bayes, Logistic Regression, and Support Vector Machines (SVM). These models are selected with an emphasis of their simplicity, efficiency and interpretability pertinent to sentiment analysis task. They are all peculiar in that they have relative strengths and weaknesses, which are analyzed and compared in the following stages.

Training:
After those models are selected, the models are then trained using the pre-processed and vectorized training data. During the training phase, the models are trained to understand alias relationships between the input features which are textual data and the output labels that are positive or negative sentiment. The training process involves adjusting the model parameters to minimize the prediction error and improve performance.

Evaluation:
After training is complete the test set is used to compare the performance of the models and their ability to generalize. Metrics like accuracy, confusion matrix and classification report about the sentiments of models are used. Moreover, precision, recall, and F1-score are computed in order to get more information concerning the performance of the models concerning the positive and negative sentiment class. Such an evaluation leaves little room for identifying the strengths and weaknesses of the models under comparison.

Project Implementation:
The project is done in Python language with necessary libraries like pandas, Scikit-learn, and nltk for data handling and manipulation, and machine learning and natural language processing techniques. The data is read into a pandas DataFrame, and the rest of the preprocessing is done using in-built functions in pandas, as well as various modules in the nltk library. Feature extraction, model selection, and training as well as the evaluation of the model is done using the scikit-learn library with an overall effective platform for machine learning.

The implementation adheres to these software engineering principles: modularity, documentation and version control to make the code accessible, replicable and sustainable. Code is grouped at the project level and divided into logical modules and functions, which facilitates its further use in additional research and development. Furthermore, the proper comments and documentation are added to give reasons for the process and help the team members.

In general, the methodology and the project implementation outline the approach that can be followed to perform the sentiment analysis on the IMDB movie reviews dataset. By following those steps, the researchers and practitioners can properly pre-process the data, choose proper models for analysis, train and test the models and make conclusions about the sentiment in textual data.

IV. RESULT

The results of the sentiment analysis experiments conducted on the IMDB movie reviews dataset using Naive Bayes, Logistic Regression, SVM, Decision Trees, and Random Forests are presented below. Each model was evaluated based on accuracy, precision, recall, and F1-score metrics to assess its performance in classifying movie reviews as positive or negative sentiments.

	Model	Accuracy	Precision
1	Naive Bayes	0.82	0.83
2	Logistic Regression	0.84	0.86
3	SVM	0.824	0.83
4	Decision Trees	0.698	0.7
5	Random Forests	0.839	0.84

V. CONCLUSION

This work provides a comparative analysis of various Machine Learning classification algorithms for the sentiment analysis task conducted on the IMDB movie reviews dataset. Some of the models applied are Naïve Bayes, Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Random Forests. All models were evaluated using accuracy and precision measurements to evaluate how well they can detect the movie review as either positive or negative sentiment.

It can be seen from the results that Logistic Regression and Random Forests have the highest accuracies of 0. 84 and 0. 839, respectively. Logistic Regression also scored the highest precision at 0. 86, whereas Random Forests gave an accuracy of 0. 84. The study shows that both Logistic Regression and Random Forests can work efficiently for the sentiment analysis task, probably because of the former's ability to learn the relationships in the data and the latter's ability to avoid overfitting, which is beneficial for unseen data.

Naive Bayes, which had an accuracy of 0. 82 and the precision of 0. 83, also performed very well. For this reason, it can be used for sentiment analysis tasks since it does not require much computational resources, and it is more interpretable in some cases.

SVM reached the average accuracy of 0. 824 and the precision of 0. 83, indicating robust performance. However, it slightly underperformed when compared with Logistic Regression and Random Forests. Another advantage of SVM is that it performs better when data is in high dimensional space and therefore can be used for text categorization in spite of obtaining slightly lower accuracy in this work.

As it can be observed, using Decision Trees yielded the lowest accuracy of 0. 698 and the precision of the results is 0. 7. The high variance of Decision Trees might have led to overfitting of the data leading to low performance on the test set. However, Decision Trees provide interpretability and simplicity, which at times is desirable in a model.

In conclusion, Logistic Regression and Random Forests emerged as the most effective models for sentiment analysis on the IMDB dataset, combining high accuracy and precision. Naive Bayes and SVM also showed commendable performance, providing viable alternatives depending on specific requirements such as computational efficiency and model interpretability. Decision Trees, while less effective in this context, still hold value for their straightforward implementation and ease of understanding. This study underscores the importance of evaluating multiple models to identify the most suitable approach for sentiment analysis tasks.

REFERENCES

1. Surnar, Avinash, and Sunil Sonawane. "Review for Twitter Sentiment Analysis Using Various Methods."IJARCETVOL 6-ISSUE 5,2017.
2. Eliacik, Alpaslan Burak, and Erdoğan Erdoğan. "Userweighted sentiment analysis for financial community on Twitter." Innovations in Information Technology (IIT), 2015 11th International Conference on. IEEE, 2015.
3. Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani|, Veselin Stoyanov.‖SemEval-2016 Task 4:Sentiment Analysis in Twitter‖, Proceedings of SemEval2016, Association for Computational Linguistics.
4. Rasika Wagh,Payal Punde.‖ Survey on Sentiment Analysis using Twitter Dataset‖ Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Xplore ISBN:978-1-5386-0965-1
5. Anchal Kathuria, Dr. Saurav Upadhyay.‖ A Novel Review of Various Sentimental Analysis Techniques‖ International Journal of Computer Science and Mobile Computing, Vol.6 Issue.4, April- 2017, pg. 17-22
6. D. M. E.-D. M. Hussein, ―A survey on sentiment analysis challenges,‖ J. King Saud Univ. - Eng. Sci., vol. 34, no. 4, 2016.
7. Liu, B. Sentiment analysis: mining opinions, sentiments, and emotions. The Cambridge University Press.2015.
8. Bilal Saberi, Saidah Saad.‖Sentiment Analysis Or Opinion Mining: A Review‖.International Journal of Advanced Science Engineering Information Technology, Vol7(2017), ISSN:2088-5334.
9. J. Bollen, H. Mao, and X. Zeng "Twitter mood predicts the stock market". Journal of Computational Science, 2(1): 1-8 2011.
10. Joshi, N. S., & Itkat, S. A. (2014). A survey on feature level sentiment analysis. International Journal of Computer Science and Information Technologies, 5(4), 5422-5425.
11. Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. "A survey on sentiment analysis methods, applications, and challenges." Artificial Intelligence Review 55.7 (2022): 5731-5780.
12. Mehta, Pooja, and Sharnil Pandya. "A review on sentiment analysis methodologies, practices and applications." International Journal of Scientific and Technology Research 9.2 (2020): 601-609.
13. Malviya S, Tiwari A, Srivastava R, Tiwari V. Machine Learning Techniques for Sentiment Analysis: A Review. sms [Internet]. 30Dec.2020 [cited 18Jun.2024];12(02):72-8. Available from: https://www.smsjournals.com/index.php/SAMRIDDHI/article/view/2078
14. Murthy, G. S. N., et al. "Text based sentiment analysis using LSTM." Int. J. Eng. Res. Tech. Res 9.05 (2020).
15. A. B. Goldberg and X. Zhu," Seeing stars when there aren't many stars: graph-based semi-supervised learning for

sentiment categorization". In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, pp. 45-52

16. R. Prabowo and M. Thelwall." Sentiment analysis: A combined approach". Journal of Informetrics , 3(2): 143-157, 2009
17. Chen and J. Xie "Online consumer review: Word-ofmouth as a new element of marketing communication mix". Management Science, 54(3): 477-491, 2008
18. M. Abdul-Mageed, M. T. Diab and M. Korayem ―Subjectivity and sentiment analysis of modern standard Arabic‖, In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2,2011.
19. Shanta Rangaswamy, Shubham Ghosh, Srishti Jha, Soodamani Ramalingam. Metadata Extraction and Classification of YouTub Videos Using Sentiment Analysis. 2016 IEEE International Carnahan Conference on Security Technology (ICCST). DOI. 10.1109/CCST.2016.7815692
20. Mauro Dragoni. NEUROSENT-PDI at SemEval-2018 Task 7: Discovering Textual RelationsWith a Neural Network Model. Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 848–852. June 5–6, 2018.
21. Sharma A, Dey S. A boosted SVM based sentiment analysis approach for online opinionated text. InProceedings of the 2013 research in adaptive and convergent systems 2013 Oct 1 (pp. 28-34). ACM.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

📱 9940 572 462 ✆ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details