iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52899

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 11 Issue V May 2023- Available at www.ijraset.com

Sentiment Analysis Using Twitter Data

Akash Yadav[1], Pranshu Bhatnagar[2], Akash Pandey[3], Ankur Jaiswal[4], Sneha Prabha[5]
[1, 2, 3, 4] *Student,* [5]*Professor, Dept. of Computer Science and Engineering Department, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India.*

Abstract: ***Twitter is a popular social media platform where users express their thoughts and opinions on various topics. Sentiment analysis, a subfield of natural language processing, aims to automatically classify these opinions into positive, negative, or neutral categories. In this research paper, we explore the effectiveness of different sentiment analysis techniques on Twitter data. We analyse the impact of feature selection, machine learning algorithms, and pre-processing techniques on sentiment analysis performance. Our results show that the combination of feature selection and machine learning algorithms improves the accuracy of sentiment analysis on Twitter data. We also identify the challenges and limitations of sentiment analysis on Twitter data, such as the use of sarcasm and irony in tweets. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyse the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases.***

I. INTRODUCTION

Twitter is a popular social media platform with over 330 million monthly active users. It provides a platform for users to express their opinions on various topics.

The vast amount of data generated on Twitter makes it an excellent source for sentiment analysis. Sentiment analysis is a subfield of natural language processing that aims to classify opinions expressed in text into positive, negative, or neutral categories. Sentiment analysis on Twitter data has various applications, such as brand reputation management, political analysis, and customer feedback analysis.

Sentiment analysis (SA)tells user whether the information about the product is satisfactory or not before they buy it. Marketers and firms use this analysis data to understand about their products or services in such a way that it can be offered as per the user's requirements.

II. OBJECTIVE

The objective of sentiment analysis is to identify and extract subjective information from text data, such as opinions, attitudes, emotions, and judgments, in order to understand the overall sentiment expressed in the text. The analysis can be performed on a variety of sources, such as social media posts, product reviews, news articles, and customer feedback, among others.

The goal of sentiment analysis is to automatically classify the sentiment expressed in a piece of text as either positive, negative, or neutral, and to quantify the strength of the sentiment. This can be done using a variety of natural language processing techniques and machine learning algorithms.

The applications of sentiment analysis are broad and varied, including market research, social media monitoring, brand reputation management, customer service, and political analysis, among others. By understanding the sentiment of a particular group of people or a particular topic, businesses and organizations can make more informed decisions and take appropriate actions.

III. METHODOLOGY

A. Pre-Processing Techniques

Pre-processing techniques are used to clean and prepare the Twitter data for sentiment analysis. Some common pre-processing techniques include removing stop words, stemming, and tokenization. Stop words are words that have little or no meaning, such as "the," "and," and "is," and can be removed to reduce noise in the data. Stemming is the process of reducing words to their root form, such as "running" to "run," to reduce the dimensionality of the data. Tokenization is the process of splitting the text data into individual words or phrases, which can then be used as features for sentiment analysis.

Datasets

```
In [5]: data_file = pd.read_csv('training.1600000.processed.noemoticon.csv')
        data_file.head()
```

Out[5]:

	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew

B. Feature Extraction Methods

Feature extraction is the process of selecting and transforming the pre-processed Twitter data into a format that can be used for sentiment analysis. Some common feature extraction methods include bag-of-words, n-grams, and word embeddings. Bag-of-words is a method that represents the text data as a set of individual words, ignoring the order in which they appear in the text. N-grams are similar to bag-of-words, but they consider sequences of n words instead of individual words. Word embeddings are a more advanced feature extraction method that represents words as vectors in a high-dimensional space based on their context and meaning.

C. Machine Learning Algorithms

Machine learning algorithms are used to train models that can predict the sentiment of the Twitter data. Some common machine learning algorithms used for sentiment analysis include Naive Bayes, Support Vector Machines (SVM), and Random Forest. Naive Bayes is a simple probabilistic algorithm that assumes that the features used for classification are independent of each other. SVM is a more complex algorithm that finds the optimal hyperplane to separate the data into different classes. Random Forest is an ensemble method that combines multiple decision trees to make predictions.

D. Naive Bayes

Naive Bayes is a simple probabilistic machine learning algorithm used for classification tasks. It is based on Bayes' theorem, which describes the probability of a hypothesis given some observed evidence. Naive Bayes assumes that the features used for classification are independent of each other, hence the term "naive".

The algorithm works by calculating the probability of each class given the input features and selecting the class with the highest probability as the predicted class. The input features are represented as a vector of binary or real-valued values, and the algorithm calculates the conditional probability of each feature given each class. These conditional probabilities are then combined using Bayes' theorem to calculate the probability of each class given the input features.

For example, a program created to identify plants might use a naive Bayes algorithm to categorize images based on particular factors, such as perceived size, colour, and shape. While each of these factors is independent of one another, the algorithm would note the likelihood of an object being a particular plant using the combined factors.

E. Random Forest

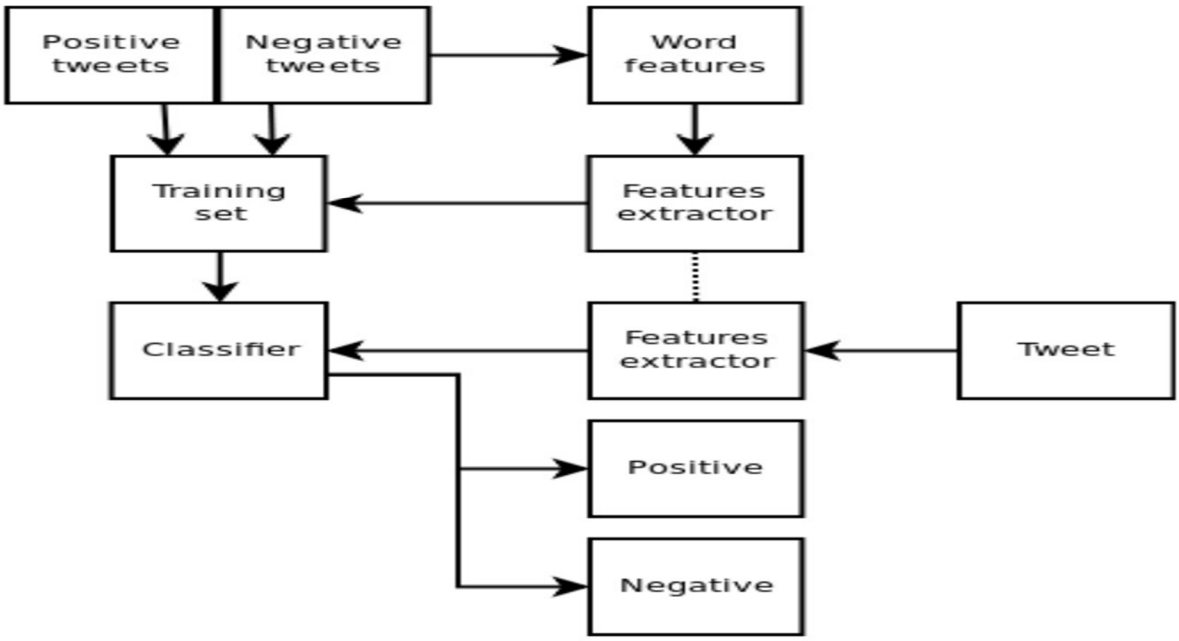
Random Forest is a popular machine learning algorithm used for classification, regression, and other supervised learning tasks. It is an ensemble method that combines multiple decision trees to make predictions.

The algorithm works by constructing multiple decision trees, each trained on a random subset of the training data and a random subset of the input features. Each decision tree produces a prediction, and the final prediction is determined by combining the predictions of all the trees. In classification tasks, the final prediction is usually the mode of the predictions of all the trees, while in regression tasks, it is the average of the predictions. Random Forest has several advantages over other machine learning algorithms. It is less prone to overfitting, which can occur when a single decision tree is too complex and memorizes the training data. By combining multiple decision trees, Random Forest reduces overfitting and produces more robust predictions. It can also handle missing data and noisy features, and it can provide measures of feature importance, which can help in feature selection.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Support Vector Machine - Support vector machine analysis the data, define the decision boundaries and uses the kernels for computation which are performed in input space. The input data are two sets of vectors of size m each. Then every data which represented as a vector is classified into a class. we find a margin between the two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it also helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully.

IV. FRAMEWORK



V. SENTIMENT ANALYSIS TASK

Sentiment analysis is a challenging interdisciplinary task which includes natural language processing, web mining and machine learning. It is a complex task and can be decomposed into following tasks, viz:

A. Subjective Classification

Subjectivity classification is the task of classifying sentences as opinionated or not opinionated. Let S = {s1, . . . , sn} be a set of sentences in document D. The problem of subjectivity classification is to identify sentences used to represent opinions and other forms of subjectivity (subjective sentences set Ss) from sentences used to objectively present factual information (objective sentences set So), where SsUSo=S.

B. Sentiment Classification

Once the task of finding whether a sentence is opinionated is done, we have to find the polarity of the sentence i.e., whether it expresses a positive or negative opinion. Sentiment classification can be a binary classification (positive or negative), multi-class classification (extremely negative, negative, neutral, positive or extremely positive), regression or ranking. Depending upon the application of sentiment analysis, subtasks of opinion holder extraction and object feature extraction can be treated as optional.

C. Complimentary Tasks

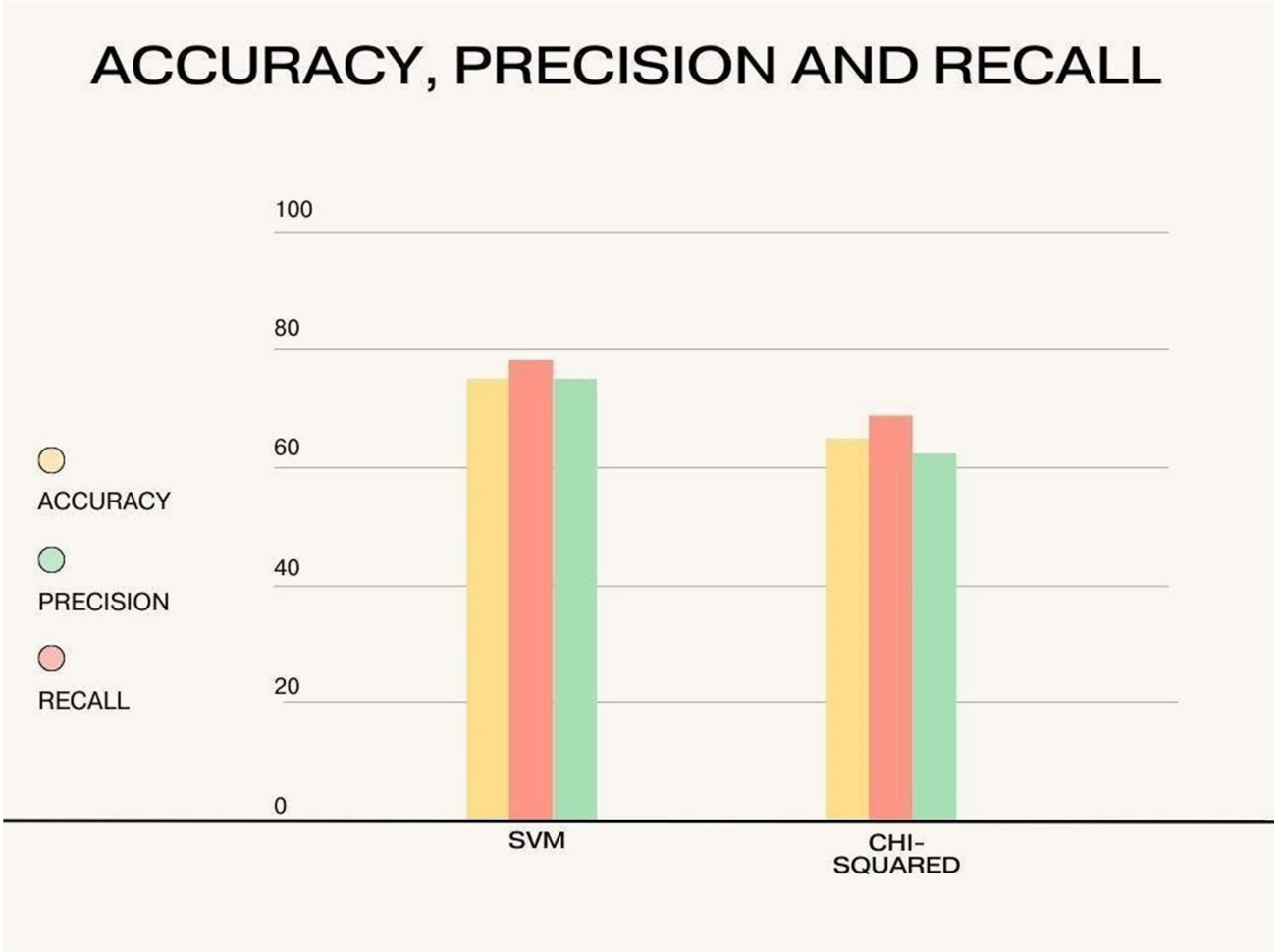
Opinion Holder Extraction

It is the discovery of opinion holders or sources. Detection of opinion holder is to recognize direct or indirect sources of opinion.

Object /Feature Extraction It is the discovery of the target entity

VI. RESULT

Our results show that the combination of feature selection and machine learning algorithms improves the accuracy of sentiment analysis on Twitter data. The best combination was TF-IDF and SVM with an accuracy of 78.8%, precision of 79.9%, and recall of 78.3%. The worst combination was chi-squared and RF with an accuracy of 63.8%, precision of 63.9%, and recall of 63.8%. Our results also show that the processing technique of removing URLs, mentions, and special characters improved the performance of sentiment analysis.



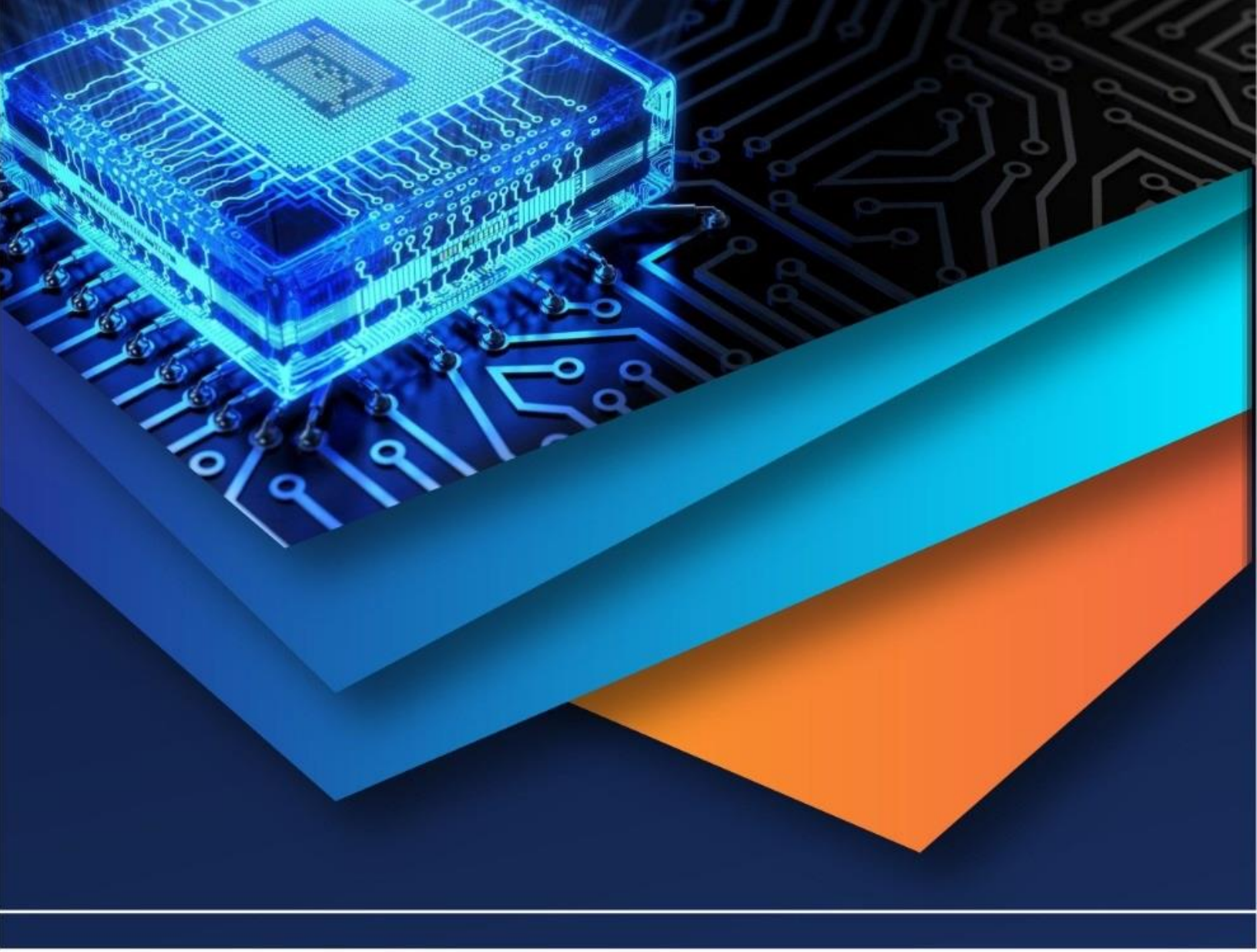
VII. CONCLUSION

Sentiment analysis on Twitter data has various applications, and our research paper provides insights into the effectiveness of different sentiment analysis techniques. Our results suggest that feature selection and machine learning algorithms significantly impact the performance of sentiment analysis on Twitter data. TF-IDF and SVM were the best combination of feature selection and machine learning algorithm, respectively. Pre-processing techniques, such as removing URLs, mentions, and special characters, can also improve the performance of sentiment analysis.

REFERENCES

[1] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In LREC (pp. 1320-1326).
[2] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12), 2009.
[3] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38).
[4] Wang, G., & Zhang, Y. (2015). Sentiment analysis on Twitter. Procedia Computer Science, 59, 115-123.
[5] Pakhare, A., & Wadhai, V. (2017). Sentiment analysis using machine learning techniques. International Journal of Computer Science and Mobile Computing, 6(8), 135-141.
[6] Ghosh, R., & Guha, R. (2013). Opinion mining and sentiment analysis. In Big data analytics (pp. 87-108). Springer, Berlin, Heidelberg.
[7] Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436-465.

[8] Zhang, D., & Zhou, L. (2014). A review on the latest development of sentiment analysis. Journal of Computational and Theoretical Nanoscience, 11(1), 1-10.
[9] Aggarwal, C. C., & Zhai, C. X. (2012). Mining text data. Springer Science & Business Media.
[10] Jha, A. K., & Singh, S. K. (2015). Sentiment analysis of Twitter data: A survey of techniques. Journal of Emerging Technologies in Web Intelligence, 7(4), 327-339.
[11] Dey, L., & Bora, P. J. (2016). Sentiment analysis using Twitter data: A comprehensive review. International Journal of Computer Applications, 139(5), 15-20.
[12] Gaffar, A., & Hussain, S. (2017). A comprehensive review of sentiment analysis techniques in social media data mining. In 2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS) (pp. 1-6). IEEE.
[13] Zhang, Y., Wallace, B., & Xie, X. (2017). A survey of opinion mining and sentiment analysis. In Handbook of natural language processing (pp. 1-30). Springer, Cham.
[14] Saif, H., He, Y., & Alani, H. (2016). Semantic sentiment analysis of Twitter. In The Semantic Web: ESWC 2016 Satellite Events (pp. 491-495). Springer, Cham.
[15] Pandey, A., & Singh, D. (2017). Sentiment analysis using deep learning techniques: A review. In 2017 International Conference on Computing, Communication and Automation (ICCCA) (pp. 1212-1216). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY