**SENTIMENT ANALYSIS OF TWITTER DATA USING TF-IDF AND MACHINE LEARNING TECHNIQUES**

1 Akshaya C, 2 Arun Kumar S, 3 Barath Kumar S, 4 Kalaiyarasan N

1Assistant Professor, Department of Computer Science Engineering, Adhiyamaan College of Engineering(Autonomous),Hosur, India

2,3,4  Student, Department of Computer Science Engineering., Adhiyamaan College of Engineering(Autonomous),Hosur, India



**Abstract*-*** Social media platforms have become integral to our daily lives over the past two decades. Nowadays, it is very important to get information from social media, keep track of trends in social media, and learn about people's feelings and emotions on social media. Sentiment analysis of a Twitter text was used in this study to investigate the subjective polarities of the writings. Positive, negative, and neutral are the polarities. A public data set has been obtained during the sentiment analysis's initial stage. Second, in order to get the data ready for machine learning training, natural language processing methods were used. The predictive results are obtained through the effective implementation of machine learning models with TF-IDF and the training and testing of a data set.

1. **INTRODUCTION**

The rise of the modern era brought with it a new mode of social interaction and communication: social media platforms. Whether we accept it or not, SM became ingrained in our daily lives at random; It became an indispensable component of our current way of life. People today use social media to share their feelings, desires, and ideas about a particular topic and market political messages, among other things. Twitter's platform receives a lot of attention from people in the previous category. It is evident that the majority of politicians worldwide are using Twitter as their no. one's favorite platforms, but we shouldn't forget that each platform has its own advantages and disadvantages that influence people's decisions. One thing that most people know about all social media platforms is that most of the time, their disadvantages outweigh their advantages. From philosophy to machine learning (ML), natural language processing (NLP) and other data analysis evaluating techniques can be used to analyze the activities on this platform. This is the right time to focus on this new phenomenon because of all of the important reasons for social media.

1. **PROPOSED SYSTEM**

The most essential processes, the data cleaning and selection level, are applied in the proposed method, which begins with the collected tweets. in text cleaning, which encompasses all of the NLP methods utilized to prepare the text for conversion. At the level of text polarity, the subjectivity of each cleaned tweet has been considered. BoW, or bag of words, has been used to convert categorical data—that is, textual data—into numerical data at the text-to-number level. Finally, the Random Forest classifier was used to train and test the data in the ML to achieve the desired outcomes.The development of artificial intelligence can benefit from this. High degree of precision in the application of machine learning algorithms. "To determine the feelings, real-time dataset is utilised."

1. **WORKING**

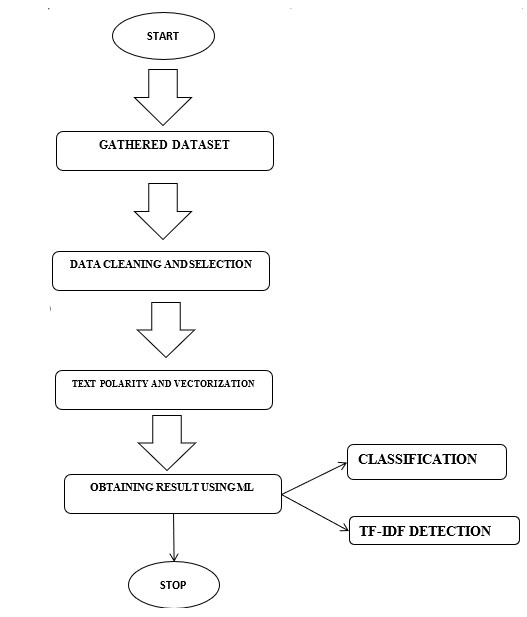
**CLASSIFICATION**

A crucial step in any supervised learning is this method. This procedure gives the agent the ability to learn from its experiences by training more than half of the data, even though the agent does not have any default information about the environment. 70% of the dataset is typically given to the agent so that it can learn from the training; The remaining thirty percent is used to evaluate the classifier's accuracy and determine whether or not it functions properly. A different classifier must be used on the hosted data if the suggested ML algorithm fails

**TF-IDF DETECTION**

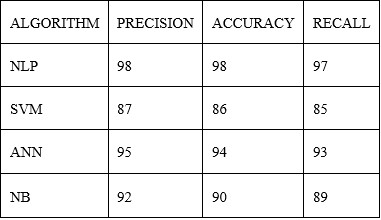
A numerical statistic called term frequency–inverse document frequency is meant to show how important a word is to a document in a collection or corpus. In user modeling, text mining, and information retrieval searches, it is frequently used as a weighting factor.

**ARCHITECTURE DIAGRAM**

****

**RESULT AND DISCUSSION**

The classifier's performance has been evaluated using the parameters and attributes of the dataset. This comparison's outcomes are shown. As can be seen from the graph, the performance of our proposed model is superior to that of the other model that is already in use. The result is presented. As can be seen from the graph, increasing the sample size enables us to improve the system's performance. This is how we explain it: NLP provides a good balance between coverage (unigrams) and the ability to capture sentiment expression patterns (trigrams). However, the improvement may not be achieved by simply increasing the size of the training data once the dataset reaches a certain size. In order to get the best possible result, we looked at two versions of our proposed model.



1. **CONCLUSION**

Machine learning and lexicon-based approaches to opinion mining, as well as cross-domain and cross-lingual methods and a few evaluation metrics, are the subjects of our survey and comparative study. Our research demonstrates that the NLP-based machine learning methods we proposed outperform other algorithms, such as SVM and naive Bayes, in terms of output. Lexicon-based methods, on the other hand, are very effective in some cases and require little effort in human-labeled documents. We also investigated the effects of various features on classifiers. We can draw the conclusion that more accurate results can be obtained with cleaner data. The sentiment accuracy of the machine learning model is superior to that of other models.

1. **REFERENCE**

[1] A. Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] R. Parikh and M. Movassate, “Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

[3] Go, R. Bhayani, L. Huang. “Twitter Sentiment Classification Using Distant Supervision". Stanford University, Technical Paper,2009

[4] L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitter from Biased and Noisy Data". COLING 2010: Poster Volume, pp. 36-44.

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[6] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011, pp. 30-38

[7] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volume pages 241{249, Beijing, August 2010

[8] Neethu M, S and Rajashree R,” Sentiment Analysis in Twitter using Machine Learning Techniques” 4th ICCCNT 2013, at Tiruchengode, India. IEEE – 31661

[9] Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets", 8th international workshop on Semantic Evaluation (Sem Eval 2014), Dublin, Ireland, Aug 23-24 2014, pp 171-175.