Twitter Sentiment Classification With Tf-Idf And Machine Learning Techniques

Beaula pinky B[1], Sangeetha V[2], Saranya.K[3], Shalini A[4]
[1]Assistant Professor, Dept of Computer Science and Engineering
[2, 3, 4]Dept of Computer Science and Engineering
[1, 2, 3, 4]Sri Shanmugha College of Engineering & Technology,
Pullipalayam, Morur (P.O), Sankari (T.K). Salem (D.T.), Tamil Nadu–637304

Abstract- *Social media platforms have become an integral part of 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 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.*

Keywords- NLP, sentimental analysis, machine learning

I. 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 not only share their feelings, desires, and ideas about a particular topic, but also to 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.

NLP

NLP is a subfield of AI and computer science, especially machine learning. It examines the human language and the computer's comprehension of it. Computational linguistics can aid in the acquisition of this method. A lot of knowledge about the lexicon, semantics, and syntax, as well as information about our real world, are required to comprehend natural language. NLP can be thought of as a combination of linguistic philosophy, computer science, and artificial intelligence. The interactions between human language and computers (Robot agent) are the focus of this scientific subfield. This field is concerned with computer programming and coding in order to process and analyze natural language data.

MACHINE LEARNING

The study of methods that "learn," or methods that use data to improve performance on a particular set of tasks, is the focus of the field of study known as machine learning (ML). It is thought to be a component of artificial intelligence.

In order to make predictions or decisions without being explicitly programmed to do so, machine learning algorithms construct a model from sample data, or training data. When it is difficult or impossible to develop conventional algorithms that can carry out the required tasks, machine learning algorithms are used in a wide range of applications, including computer vision, agriculture, speech recognition, email filtering, and medicine.

SENTIMENTAL ANALYSIS

Natural language processing, text analysis, computational linguistics, biometrics, and emotion AI are all used to systematically identify, extract, quantify, and study affective states and subjective information in sentiment analysis (also known as opinion mining or emotion AI). In a variety of fields, including marketing, customer service, and clinical medicine, sentiment analysis is frequently used with online and social media voice of the customer materials like reviews and survey responses. With the rise of deep language

models like, it is now possible to analyze more difficult data domains like news texts, where authors typically express their opinions and feelings less explicitly.

OBJECTIVES

To get the precise sentiments of each tweet in the taken sample dataset, we used natural language processing (nlp) and machine learning (TF-IDF) in our project to detect sentimental analysis in Twitter using the input dataset.

II. RELATED WORKS

According to PRATHAMESH S et al., social media is a platform where individuals are free to share their thoughts, address issues, and voice their opinions. Before getting into the details, it's important for people to understand what social media is. People can share or exchange information, images, videos, ideas, and many other things with one another through a particular networking medium on a digital interactive platform known as social media. According to Ludwig Wolff et al., it has also become increasingly simple to access a significant portion of the data shared on social media platforms over time. Behavioral economics, which challenges conventional notions of efficient markets and examines the rationality of consumers in the market, is quickly emerging as an intriguing new viewpoint for investors. According to Alexander Pak et al., microblogging is now a very common means of communication among Internet users. Every day, millions of users voice their thoughts on various aspects of life. According to Sayali P. Nazare's proposal, sentiment analysis is the process of categorizing and identifying opinions or feelings that are expressed in source text. Today, microblogging is a very common method of communication among Internet users.

III. PROPOSED METHOD

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 (natural language processing) 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.

"This can be used in the development of artificial intelligence." "High level of accuracy in the use of machine learning techniques." "Real time dataset is used to identify the sentiments."

GATHERED DATASET

The kaggle online community for data scientists and ML practitioners provided the used dataset, which had already been prepared.

DATA CLEANING AND SELECTION

Because manipulating a dataset necessitates specialized commands, data analysis should be performed on all datasets. Importing a dataset, performing the majority of actions on its columns and rows, appending and deleting records, and other actions are all performed by data analysis on any dataset. Applying NLP and ML algorithms would be impossible without data analysis. In any study conducted in the same field, this step will determine which features should be used and which should be eliminated.

TEXT CLEANING

Before using any classifier algorithms, the data—in our case, Twitter texts—must be thoroughly cleaned and prepared. There are numerous elements (such as hashtags, emoticons, unconventional punctuation, spaces, and symbols) that cannot be classified and must be eliminated (filtered out) from every text. This step saves storage space by compressing our data, which is one of its greatest benefits. The performance of the work can be improved by reducing the size of the hosted dataset, and the data size can be used for informational purposes. The result section contains specifics. The following text cleaning procedures, which include the following steps, have been applied to the dataset in the experimental portion: First, stopwords were removed; then, word lemmatizing was used to change the words into their roots; and finally, regular expressions were used to get rid of links, emails, and other things.

TEXT POLARITY AND VECTORIZATION

One of the study's main objectives is this section. Preparing the text for subjective sentiment polarities (or, in some resources, sentiment score) is what we do from the beginning to the end. A technique for determining the subjectivity of each tweet is text polarity. The tweets are not objective because the subject is a human and is tweeting his own thoughts about a particular event or anything else. To be divided into the three levels of positive, negative, and neutral, it must be discovered. Each sentence in our experimental work has been evaluated after being cleaned using NLP techniques.

Each textual data (in this instance, a tweet from Twitter) has three possible labels: positive, negative, or neutral In this study, we first adapted the following measurement to determine the sentiment polarity of each tweet.

Sentiment score = positive - negative / positive + negative + 2

SPLITTING INTO TRAIN AND TEST

A technique that is a kind of division is necessary for each and every machine learning algorithm. The entire dataset will be divided into two parts during this procedure: Testing and instruction. The researcher decides how many percentages to divide each part into. It can be 80% for testing and 20% for training, respectively. likewise 30% and 70%. The performance of a machine learning algorithm can be evaluated using this approach. As previously stated, the dataset must be divided into two subsets for this procedure:

Set of training: used for the machine learning model's fitting and training.

Set of tests: utilized to evaluate the machine learning model's fit.

IV. OBTAINING RESULT USING ML

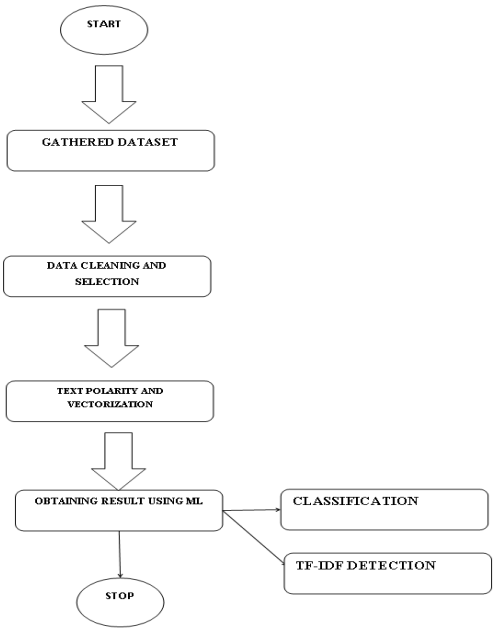
CLASSIFICATION

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. The splitting process in ML will be described in detail in the figure below**.**

TF-IDF DETECTION

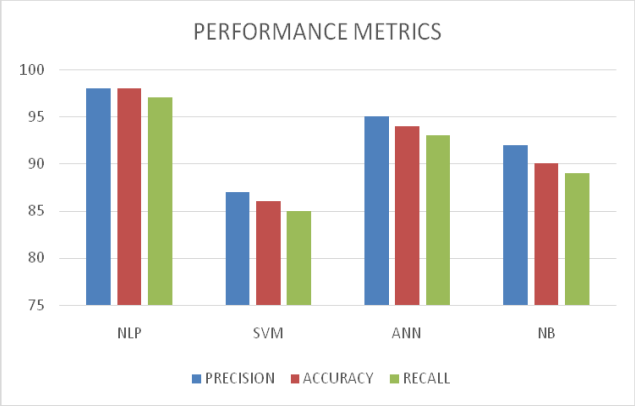
A numerical statistic called term frequency–inverse document frequency is meant to show how important a word is to a document in a collection or corpus. In user modeling, text mining, and information retrieval searches, it is frequently used as a weighting factor. In order to account for the fact that some words appear more frequently in general, the tf–idf value increases in proportion to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word. One of the most widely used term-weighting methods in use today is tf–idf.

V. ARCHITECTURE DIAGRAM



VI. RESULT AND DISCUSSION

The classifier's performance has been evaluated using the parametes 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.



ALGORITHM	PRECISION	ACCURACY	RECALL
NLP	98	98	97
SVM	87	86	85
ANN	95	94	93
NB	92	90	89

VII. 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. In order to improve sentiment classification accuracy and adaptability to a variety of domains and languages, we can concentrate on the study of combining the opinion lexicon method with machine learning.

REFERENCES

[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- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

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

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

[5] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference 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 Volumepages 241{249, Beijing, August 2010

[8] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

[9] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

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

[11] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[12] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[13] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[14] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013, DOI: 10.1007/s11280-013-0268-7.

[15] Liu, S., Li, F., Li, F., Cheng, X., &Shen, H.. Adaptive cotraining SVM for sentiment classification on tweets. In Proceedings of the 22nd ACMinternational conference on Conference on information &knowledgemanagement (pp. 2079-2088). ACM,2013.

[16] Pan S J, Ni X, Sun J T, et al. "Cross-domain sentiment classification viaspectral feature alignment". Proceedings of the 19th internationalconference on World wide web. ACM, 2010: 751-760.

[17] Wan, X..“A Comparative Study of Cross-Lingual SentimentClassification”. In Proceedings of the The 2012 IEEE/WIC/ACMInternational Joint Conferences on Web Intelligence and IntelligentAgent Technology-Volume 01 (pp. 24-31).IEEE Computer Society.2012

[18] Socher, Richard, et al. "Recursive deep models for semanticcompositionality over a sentiment Treebank." Proceedings of theConference on Empirical Methods in Natural Language Processing(EMNLP). 2013.

[19] Meng, Xinfan, et al. "Cross-lingual mixture model for sentimentclassification." Proceedings of the 50th Annual Meeting of theAssociation for Computational Linguistics Volume 1,2012

[20] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., &Stede, M..“Lexiconbasedmethods for sentiment analysis”. Computational linguistics, 2011:37(2), 267-307.