Sentiment Analysis of Twitter Tweets

Vijay Kumar Gumasa[1], Thota Rakesh Kumar[2], P. Rahul Das[3]

[1,2,3]Computer Science and Engineering GCET Hyderabad, Telangana, India

Received: 18.04.2024 Revised : 16.05.2024 Accepted: 24.05.2024

ABSTRACT
Social networking is becoming more and more popular these days. Public and private opinion is constantly expressed and spread on a wide range of topics through various social media platforms. Among the social media platforms that are becoming more and more popular is Twitter. Twitter provides businesses with an efficient and quick method to examine consumer viewpoints about the essential elements of market success. One method to measure consumers' perceptions computationally is to create a sentiment analysis program. This research presents a sentiment analysis design that extracts a large number of tweets. In this development, prototyping is employed. The findings, which are displayed in an HTML website and pie chart, divide the opinions of customers who tweeted into two categories: favorable and negative. Though Django is designed to run on a Linux server or LAMP, there are limitations that prevent the program from developing on a web application system. Therefore, in order to move further, this strategy must be used.

Keywords: Component Twitter, sentiment, opinion mining, social media, natural language processing.

1. INTRODUCTION

There is still a lot of opportunity for more research in the field of sentiment analysis in micro blogging because it is a relatively new topic of study. Regarding sentiment analysis of user evaluations, papers, web blogs/articles, and general phrase level sentiment analysis, a respectable amount of relevant earlier work has been done. These vary from Twitter primarily in that users are forced to convey their opinions in extremely brief words due to the 140 character tweet constraint. Naive Bayes and Support Vector Machines, two supervised learning approaches that produce the greatest results in sentiment classification, are exceedingly costly to manually label. Scientists experimenting with novel characteristics and classification methods frequently simply contrast their findings with the baseline performance. The best features and classification strategies for a given application must be selected by conducting formal comparisons of the findings obtained through various features. Social media platforms, which represent the combined ideas, beliefs, and feelings of millions of users, have developed into informational gold mines in today's hyper connected society. Because Twitter is real-time and has a large user base, it offers a wealth of source of textual data that can be harnessed for various analytical purposes. Sentiment analysis, also known as opinion mining, has emerged as a crucial research area, aiming to understand the sentiment or emotional tone expressed in textual data. This mini project focuses on harnessing the power of sentiment analysis specifically on tweets from Twitter. By analysing tweets, we aim to unravel the underlying sentiments of users, thereby gaining insights into public opinion on diverse topics, products, events, or trends.Among all the different kinds of since text messaging allows users to communicate their thoughts and feelings on a wide range of issues, it is regarded as one of the most noticeable modes of communication. The act of examining and turning unstructured text data into structured data in order to derive valuable insights is known as text mining. It is described as a versatile research approach that may be used to objectively and methodically discover features of huge sample data in order to examine a variety of topics.Text mining is a branch of data mining that builds on traditional data mining techniques. It is useful for creating complex formulations through the use of text classification and clustering algorithms (Yang, et al.). The length of the tweet, the frequent use of acronyms, misspelling words and acronyms, transliterating non-English terms using Roman scripts, confusing semantics, and synonyms are only a few of the major issues identified by Hossny et al. for text analysis on Twitter. A variety of social media platforms, including blogs, review social networks, and Twitter, are used to analyze information in order to get user opinions about certain brands, companies, or circumstances. Sentiment analysis, which measures an individual's behavior, has an important component that is composed of the attitude and feelings. A field of study called sentiment analysis (SA) can be used to further examine these sentiments (Singh, et al. SA is a subfield of natural language processing (NLP) (Chen, et al.), a cognitive computing study of people's views, sentiments, emotions, assessments, and attitudes toward entities such as goods,

services, organizations, people, problems, events, topics, and their attributes. NLP has been actively researching SA. Additionally, it seeks to evaluate and glean knowledge from the arbitrary data posted online (Basiri, et al. Sentiment analysis of user-generated data is a highly helpful tool for understanding the general consensus. The literature describes two primary methods for sentiment analysis of text documents: machine learning-based methods andapproaches based on symbolic techniques. Symbolic approaches make use of lexicons and other language tools to ascertain the tone of a particular text. Machine learning has been used in some research to classify the sentiment of a given text. This approach sometimes takes the most symbolic techniques and looks for positive, negative, and neutral categories; other times, however, it also takes into account other sentiment categories like anger, joy, and sadness.

2. RELATED WORK

The current Twitter sentiment analysis systems use a combination of natural language processing and machine learning algorithms to sort tweets into positive, negative or neutral sentiments. These systems typically feature data collection, pre-processing, feature extraction, sentiment classification, and result interpretation components. Twitter collects data by utilizing APIs or accessing online datasets. Toxic noise is managed through pre-processing techniques like tokenization, stop word removal, and stemming. Feature extraction techniques extract relevant linguistic and semantic features from tweets, which then feed into machine learning models for sentiment classification. Using algorithms like Naive Bayes, SVM, or deep learning architecture such as Recurrent Neural Networks (RNNs) and Transformers these models are trained on given datasets. The outcomes are analyzed using metrics such as accuracy, precision, recall, and F1-score, which can help to gauge public sentiment.Most sentiment analysis researchers have employed supervised machine learning algorithms, like Chauhan et al.'s, for primary categorization. Moreover, a large number of recent studies (Al-Laith et al., Yadav et al.) employ Twitter as their main data source. D-RVFL, a non-iterative deep random vectorial functional link, was utilized by Henríquez and Ruz. They examined two distinct datasets. Ten thousand tweets from the 2017 Catalan referendum are collected in dataset 1, and twenty-one hundred seventy-seven tweets from the 2010 Chilean earthquake are collected in dataset 2. They treat the datasets as though they were a two-class classification issue with positive and negative labels.

The results indicate that D-RVFL performs better than SVM, random forest, and RVFL.Ankit and Saleena presented the idea of an ensemble classification system made up of various classifiers, including logistic regression, random forest, SVMs, and naive Bayes. Their solution makes use of two algorithms: one determines the tweet's positive and negative scores, and the other uses these scores to infer the sentiment of the tweet. Additionally, there are 56,457 good and 43,532 negative tweets in the sample. The preprocessing methods were assessed by Symeonidis et al. based on the quantity of features generated and the classification accuracy that was achieved.While the detection accuracy is low, this paper focused on lemmatization, number removal, and contraction replacement strategies. They employed two datasets with the classes of positive, negative, and neutral to test four classification algorithms: logistic regression, Bernoulli Naive Bayes, linear SVC, and convolutional neural networks. Sailunaz and Alhajj evaluated the influence scores of users based on multiple user- and tweet-based criteria and employed a dataset to extract sentiment and emotion from tweets and their replies.The study presents the agreement score, sentiment score, and emotion score of replies in the influence score computation. The dataset also includes replies to tweets. Ruz and colleagues examined five classifiers and evaluated their results using two Twitter datasets containing two distinct important events. Based on Spanish datasets, they came to the conclusion that the behavior of random forests and support vector machines (SVMs) is the same in both languages.They use a Bayes factor strategy to automatically regulate the amount of edges supported by the training examples in the Bayesian network classifier, producing more

Realistic networks. Wang et al. presented a system for tracking sentiment in the general public from a social media stream (Twitter) using enhanced latent Dirichlet allocation (LDA) technique for sentiment classification and extensive multilevel filters. Using real-world content, they were able to achieve a 68% accuracy rate for general sentiment analysis. Additionally, they employed two datasets: one with four categories (junk), the other with three categories (positive, negative, and neutral). Ali et al. used machine learning and deep learning models to classify sentiment and emotions using bilingual (English and Urdu) data from Twitter and news websites.In order to find out how people are feeling about the COVID-19 pandemic, Kaur and Sharma used API to gather tweets regarding the virus that were useful. They then divided the tweets into three categories: neutral, negative, and positive. Nuser et al. presented an unsupervised learning framework for sentiment analysis on a binomial dataset gathered from Twitter, based on a serial ensemble of various hierarchical clustering techniques. Machuca et al. performed sentiment analysis on English COVID-19 pandemic tweets by utilizing a logistic regression technique on a binomial dataset that had labels for both positive and negative outcomes.

3. PROPOSED SYSTEM

For Sentiment Classification Of Twitter Tweets, A Dl Technique Of Gated Attention Recurrent Network (Garn) Is Proposed. Te Twitter Dataset (Sentiment140 Dataset) With Sentiment Tweets That The Public Can Access Is Initially Collected And Given As Input. After Collecting Data, The Next Stage Is Pre-Processing The Tweets. In The Pre-Processing Stage, Tokenization, Stopwords Removal, Stemming, Slang and Acronym Correction, Removal Of Numbers, Punctuations &Symbol Removal, Removal Of Uppercase And Replacing With Lowercase, Character &Url, Hash tag& User Mention Removal Are Done. Now the Pre-Processed Dataset Act as Input forthe Next Process.Text Is Divided Into Classes Using A Machine Learning-Based Approach That Applies Classification Techniques. Machine Learning Approaches Can Be Broadly Divided Into Two Categories.

3.1.1. Unsupervised learning

It does not consist of a category and they do not provide with the correct targets at all and therefore rely on clustering.

3.1.2. Supervised learning

Because it is based on a labeled dataset, the model receives the labels as it goes along. When these labeled datasets are used in decision-making, they are trained to produce relevant results.

The selection and extraction of the particular collection of features needed to detect sentiment is what ultimately determines the effectiveness of each of these learning techniques. The supervised classification category mostly encompasses the machine learning technique used in sentiment analysis. In a machine learning techniques, two sets of data are needed:

1. Training Set
2. Test Set.

Several machine learning methods have been developed to categorize the tweets into groups. Sentiment analysis has greatly benefited from the application of machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM). Gathering training datasets is the first step in machine learning. Using the training data, we then train a classifier. Feature selection is a crucial choice to make after a supervised classification method has been chosen. They can tell us how documents are represented.

The most commonly used features in sentiment classification are:

- Term presence and their frequency
- Part of speech information
- Negations
- Opinion words and phrases

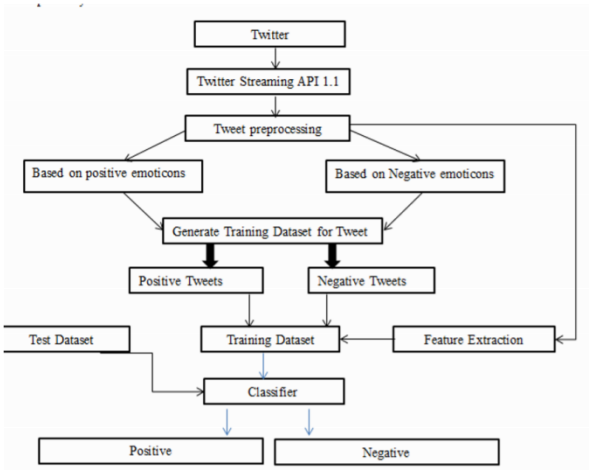

Fig 1. Sentiment Classification Based On Emoticons

Support vector machines (SVM), Naive Bayes, and Maximum Entropy are three of the most often used supervised algorithms. In contrast, when it is not feasible to have an initial collection of labeled documents or opinions to classify the remaining items, semi-supervised and unsupervised procedures are suggested.

3.2 Lexicon Based Approaches

The lexicon-based approach compares opinion terms in a sentiment dictionary with the data to ascertain polarity. To describe how Positive, Negative, and Objective the words in the dictionary are, they give sentiment scores to the opinion terms. Lexicon-based techniques mostly depend on a sentiment lexicon, which is a precompiled list of well-known sentiment terms, phrases, and even idioms created for conventional communication genres, such the Opinion Finder lexicon; There are Two sub classifications for this approach:

3.2.1. Dictionary-based

It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is Word Net, which is used to develop a thesaurus called SentiWordNet. Drawback:Can't deal with domain and context specific orientations.

3.2.2. Corpus-Based

The corpus-based approach have objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques.

- Methods based on statistics: Latent Semantic Analysis (LSA).
- Methods based on semantic such as the use of synonyms and antonyms or relationships from thesaurus like Word Net may also represent an interesting solution. According to the performance measures like precision and recall, we provide a comparative study of existing techniques for opinion mining, including machine learning, lexicon-based approaches, cross domain and cross-lingual approaches, etc., as shown in Table 2

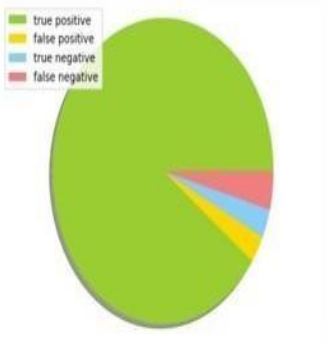

Fig 7.1 Decision Tree

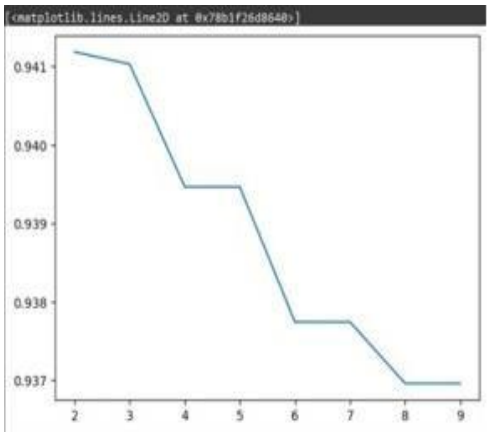

Fig 7.2 Random Forest

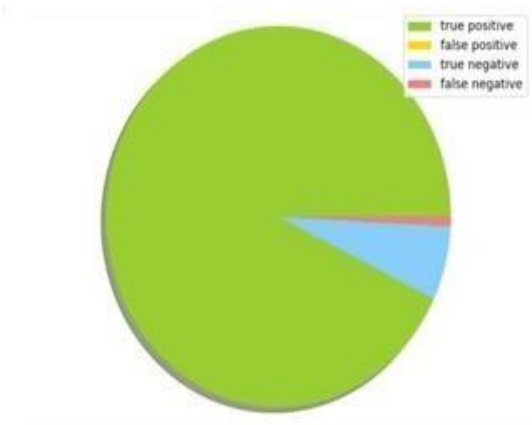

Fig 7.3 K –Nearest Neighbor's

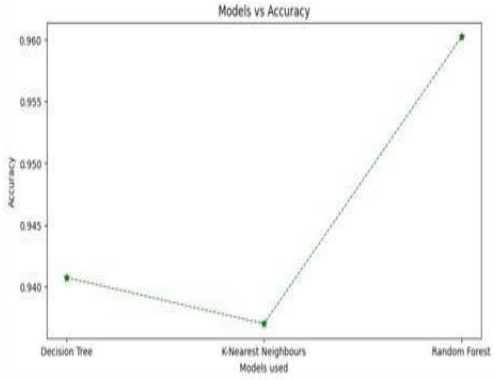

Fig.7.4 Accuracy vs.Algorithm

4. CONCLUSION

The project on "Sentiment Analysis of Tweets from Twitter" has successfullydemonstrated the implementation of a sentiment analysis system capable of analyzingpublic sentiment expressed in tweets. Through the integration of various modules,including data collection, pre-processing, feature extraction, sentiment classification,real-time processing, and user interface, the project provides valuable insights into usersentimentsonTwitter.Thesystem'saccuracyinclassifyingtweetsaspositive,negative,or neutral, coupled with its ethical considerations and user-friendly interface, makes it apowerfultoolforsocialmedia analytics. Throughout the project, several challenges were addressed, such as data noise,real-timeprocessingcomplexities, andethicalconsiderationsrelatedtouserprivacyandcontent filtering. By leveraging appropriate technologies and methodologies, thesechallenges were overcome, leading to the successful implementation of the sentimentanalysissystem. We offer an overview and comparative analysis of the state-of-the-art opinion mining techniques, including lexicon-based and machine learning methods, as well as cross-domain and cross-lingual approaches and a few assessment criteria. While lexicon-based methods are very effective in some circumstances and involve little work in human-labeled documents, research results reveal that machine learning methods, such as SVM and naive Bayes, have the highest accuracy and can be considered the baseline learning methods. The impact of different features on the classifier was also investigated. We can draw the conclusion that results can be achieved with greater accuracy the cleaner the data. When compared to other models, the bigram model offers superior sentiment accuracy. To enhance the precision of sentiment categorization and enhance its adaptability to many domains and languages, we can concentrate on investigating the amalgamation of machine learning techniques with opinion lexicon methods.

REFERENCES

[1] M.Rambocas,andJ.Gama,“Marketing Research: The Role ofSentiment Analysis”.The5thSNA-KDDWorkshop’11.University ofPorto,2013.

[2] A. K. Jose, N. Bhatia, and S. Krishna, “Twitter Sentiment Analysis”.National Institute ofTechnology Calicut,2010.

[3] P.Lai,“Extracting Strong Sentiment Trend from Twitter”.StanfordUniversity,2012.

[4] Y. Zhou, and Y. Fan, “ A Sociolinguistic Study of American Slang,” Theory and Practice in Language Studies, 3(12), 2209–2213, 2013. doi:10.4304/tpls.3.12.2209-2213

[5] Ali MZ, Javed K, Tariq A (2021) Sentiment and emotion classification of epidemic related bilingual data from social media. arXiv preprint arXiv:2105.01468

[6] Al-Laith A, Shahbaz M, Alaskar HF, Rehmat A (2021) Arasencorpus: a semi-supervised approach for sentiment annotation of a large arabic text corpus. ApplSci 11(5):2434

[7] Ankit, Saleena N (2018) An ensemble classification system for Twitter sentiment analysis. ProcediaComputSci 132:937–946. https://doi.org/10.1016/j.procs.2018.05.109

[8] Basiri ME, Nemati S, Abdar M, Cambria E, Rajendra AU, (2021) ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. FuturGenerComputSyst 115:279–294. https://doi.org/10.1016/j.future.2020.08.005

[9] Bhatnagar S, Choubey N (2021) Making sense of tweets using sentiment analysis on closely related topics. SocNetw Anal Min 11:44. https://doi.org/10.1007/s13278-021-00752-0

[10] Chauhan UA, Afzal MT, Shahid A, Abdar M, Basiri ME, Zhou X (2020) A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews. World Wide Web 23(3):1811–1829
[11] Chen J, Hossain MS, Zhang H (2020) Analyzing the sentiment correlation between regular tweets and retweets. SocNetw Anal Min 10:13. https://doi.org/10.1007/s13278-020-0624-4
[12] Cui R, Agrawal G, Ramnath R (2020) Tweets can tell: activity recognition using hybrid gated recurrent neural networks. SocNetw Anal Min 10:16. https://doi.org/10.1007/s13278-020-0628-0
[13] Dai Y, Liu J, Zhang J, Fu H, Xu Z, (2021) Unsupervised Sentiment Analysis by Transferring Multi-source Knowledge. CognComput. https://doi.org/10.1007/s12559-020-09792-8
[14] Desai M, Mehta MA (2016) Techniques for sentiment analysis of Twitter data: A comprehensive survey. In: 2016 International Conference on Computing, Communication and Automation (ICCCA). 149–154 https://doi.org/10.1109/CCAA.2016.7813707
[15] Dietterich TG, (2000) Ensemble methods in machine learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science. 1857, 1–15. https://doi.org/10.1007/3-540-45014-9_1
[16] Fatehi N, Shahhoseini HS, Wei J, Chang CT (2022) An automata algorithm for generating trusted graphs in online social networks. Appl Soft Comput 118:108475. https://doi.org/10.1016/j.asoc.2022.108475