International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-5, September 2022
https://doi.org/10.55524/ijircst.2022.10.5.7
Article ID IRP1301, Pages 50-53
www.ijircst.org

Sentimental Analysis – Detecting Tweets on Twitter

Milisha[1], Aman Jatain[2], and Priyanka Makkar[3]

[1]Student, Department of Computer Science & Engineering, Amity School of Engineering and Technology, Gurugram, India
[2]Associate Professor, Department of Computer Science & Engineering, Amity School of Engineering and Technology, Gurugram, India
[3]Assistant Professor, Department of Computer Science & Engineering, Amity School of Engineering and Technology, Gurugram, India

ABSTRACT- As we all know social media is a growing industry in the current world. People of every age are using social media directly or indirectly. Millions of people are share their thoughts on Twitter day by day. Every tweet has its own characteristics and expressions. The technologies I have used for analyzing the datasets of Twitter are data mining and NLP with Python. After collecting the data, we have trained it and made the tweets capable of testing, so it can give us the proper sentimental output. This paper will help us to understand the sentiment analysis techniques and also helps us to extract sentiments from Twitter datasets. The Twitter datasets collected from Kaggle and other sources. In this paper, we have focused on the comparative study of the different algorithms as well as on techniques.

KEYWORDS- Twitter Data, Sentimental Analysis, NLP & Mining, Naive-Bayes, Python.

I. INTRODUCTION

Every social media platform is growing dramatically irrespective of its usage and services.
Social media comes into various forms which includes online gaming platforms, dating apps, forums, online news services & social networking. Opinion transmission such as Twitter or Facebook, Business connections like LinkedIn and image sharing platforms like Instagram and so on are all
goals of different social networks. All the social networking site have one thing in alike though they want to bring people together. The influence of social
networking is so enormous that almost 1/3rd of the world's population is using it.[1]
Every platform has its own positive & negative responses on the specific person's perspective. Sometimes the impact of the speech can be spread worldwide and affect the thoughts of the population. So, we need to ensure that that impact should be positive and try to reduce negative thoughts.[3]
Some of the various number of existing social networks Twitter is currently one of the most popular, trendy and one of the majority admired data sources for researchers. Twitter is a well-known real time public microblogging network where news very often appears before it come in view in traditional news outlets. With an average of 600 million tweets every day, it has quickly grown in popularity, especially during events, thanks to its short message restriction (currently 281 characters) and unfiltered feed.[2]

II. SENTIMENT ANALYSIS

Opinion mining, also known as sentiment analysis in text mining, is a field of study which make clear of people's thoughts on any likely topic, about any likely event, and so on. It creates a large problem area. Sentiment analysis, opinion mining, opinion extraction, sentiment mining, influence analysis, subjectivity analysis, survey mining, and so on, all have different names and tasks.[6]

A. Degree of Analysis

In general, categorize of sentiment analysis can be done mainly in three different degree:

- *Document Degree Analysis*

This level determines whether the entire document conveys a +ve or -ve sentiment. This document is based on a single topic. As a result, texts that include comparative learning cannot be classified as documents.

- *Sentence Degree Analysis*

This level's job is to go through each statement and determine if it expresses a +ve, -ve, or neutral opinion. If a sentence does not express any opinion, it is considered neutral. Subjectivity categorization is linked to sentence level analysis. That expresses verifiable facts from subjective aspects and opinions, i.e. good-bad phrases, from sentences.

- *Entity/Aspect Degree Analysis*

People's preferences are not discovered through document and sentence analysis. The Entity Aspect level provides a comprehensive analysis. Previously, the Entity Aspect level was referred to as the feature level. The primary purpose of entity level is to identify constructions, whereas aspect level focuses solely on opinion or mood. It is founded on the idea

that an opinion is made up of an attitude and a final destination.

B. Related work

In the last couple decades, there has been a fast increase in sentiment analysis from user-provided text data. The classification of the text emotion processes in terms of polarity +ve, -ve, or neutral, as well as fine-grained emotions like happy, sad, boredom, furious, love, surprise, fun, neutral, enthusiasm, empty, relief, hatred, and concern, has received a lot of attention and effort. ML algorithms or a lexicon-based approach are used in this procedure.

Akshay Amolik et al. gave the idea of sentiment analysis, and they used Feature-Vector and classifiers like Nave-Bayesian and SVM to accurately classify tweets. Compared to SVM, Nave Bayesian has superior precision with the exception of lesser recall and accuracy. When it comes to accuracy, SVM performs better. The accuracy of categorization will improve as the amount of training data grows.[9]

Xing Fang and Justin Zhan obtained data from amazon.com for Sentiment Analysis on Amazon Online item Review. They used Machine Learning Algorithms to solve the complex problem of categorical sentiment polarity. They used a variety of libraries, including Naive Bayesian, SVM, and Random Forest.[11]

Prabhsimran Singh et al. zeroed in on the public authority strategy of "Demonetization" according to the viewpoint of conventional individuals by utilizing sentiment investigation with Twitter information, which was then applied to the circle of governmental issues. State-by-state examination is performed (geo-area).The sentiment analysis API utilizes a significance cloud, with six categories: "neutral, sad, very sad, very happy, and no data."[8]

Geetika Gautam and Divakar Yadav both contribute to the sentiment analysis for the classification of customer feedback. This work cab make use of Twitter data that has already been classified. In this research, they employed three supervised techniques to calculate similarity between : naive-Bayes, Max-entropy, and SVM, followed by semantic analysis, which was used in conjunction with all three algorithms. They trained and classified the following models using Python and NLTK: naive-Bayes, Max-entropy, and SVM. The Naive-Bayes technique outperforms the Max-entropy approach, whereas SVM with the unigram model outperforms SVM alone. When the Word Net of semantic analysis is put in the application after the above technique, the correctness is raised.[12]

C. Procedure

The tools that are Available for Sentiment Analysis are:

- *Tweepy*

The Twitter API is accessed using Tweepy. It's essentially a Python module that's fantastic for creating automations and Twitter bots. Tweepy's Stream Listener object keeps track of tweets continuously and captures them.[4]

- *Textblob*

Textblob is a strong Python NLP (Natural Language Processing) package. It depends on the NLTK (Natural Language Toolkit) and can be utilized for a variety of tasks, including sentence analysis, part-of-speech tagging, text hierarchies, and language interpretation.[4]

- *Pandas*

In Python, it's one from the many data frames. Pandas are ground-breaking, expressive, and adaptive information structures that simplify data analysis and control, among other things.

D. Twitter

1) Data Collection

Data collection is a crucial part of sentiment analysis. For data collection, many data sources such as online posts, blogs, microblogging sites and review sites such as Twitter and Facebook are used. For the data collection process, we used Twitter.

2) Data Preprocessing

For initiating the process we need to go through the process of collecting the data using following steps

- Stemming- In this step, we eliminate the postfix from terms such as "ing," "tion," and so on.
- Tokenization is important for Data Preprocessing because it involves substeps like "Removal of Additional Spaces," "Emoticons (C) used modified with their that actual meaning like Happy, Lonely by using Emoticon data set available over the Internet," and "Pragmatics handling like happy as happy or lub as love etc.
- Stop Word Removal- In this step, we eliminate stop words such as prepositions (a, an) and conjunctions (and, between) that aren't useful in the analysis.

3) Feature Extraction

Feature extraction determines the type of qualities that are used in opinion mining. There are several different sorts of features that are used, such as There are several different sorts of features that are used, such as

- Frequency of Terms- The frequency of any term in a text is important.[5]
- Term Co-occurrence- When a term, including such Unigram, Bigram, or n-gram, appears frequently in a sentence.
- Counting the number of verbs, adjectives, and nouns in each tweet is one of our features. Sentiment Analysis & Polarity Classification:

In all aspects of human existence, opinions, emotions, and sentiments play a significant influence. Sentiment analysis is the process of analyzing such opinions. Sentiment analysis and polarity classification are difficult tasks to do. We used a dictionary-based technique to conduct sentiment analysis." A predetermined lexicon of positive and negative words is used in this method. Most academics today utilize SentiWord net as a common lexicon for sentiment analysis.

The task of polarity classification is categorizing the reviews based on the emotions stated as Negative, Positive, or Neutral.

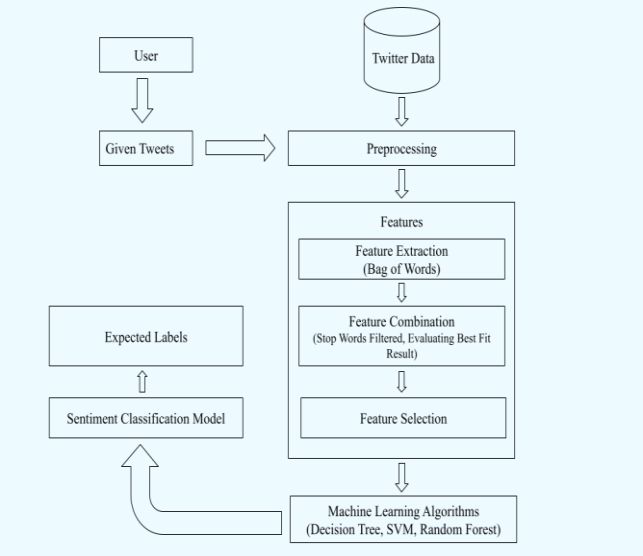

Figure 1: Architecture

This flowchart describes the suggested sentiment classification methodology's process flow in detail. The processing is represented in figure 1 by the architecture.

Training Data: - Collecting user-generated datasets (tweets). We've gathered 39,000 datasets for training purposes from GitHub, Kaggle, and other sources.

Preprocessing: The compulsion is to execute the pre-processing of data before investigating it in order to enhance the performance of analysis. To begin, the tokenize (dividing of a string's sequence) process is started, and the input stream is turned into individual words. Typically, input texts contain non-sentimental elements such as "URL's username, hashtag, punctuation, and additional white space." Simply It may be stated that using the NLTK stop word corpus, stop words from datasets that do not express any emotion can be eliminated.
All capital letters are being converted to lowercase letters. Following that, a filtration process is carried out, followed by POS tagging in order to classify words according to their parts of speech.

Features: In order to create a sentiment analysis model, we must extract each and every feature from the text input, which may be divided into morphological features and word N-gram features. We employ morphological traits to determine the presence of enlarg words (such dull, foooooool, byeeeeee, and so on), time & date expression, and punctuation signs. punctuation marks, Elongated words, and completely and partially capitalized tokens are all counted as well.[10]
Features in the Bag of Words This entails using machine learning methods to extract features from text data for modeling purposes. "It's a text representation that describes the frequency or occurrence of terms within a document." There are two items that can be used and offered as

- recognized words terminology.
- A measure of the presence of known words.

SVM is employed as a classifier in the Sentiment Analysis experiment model, and this classifier was trained over the training data.
After the 1st process was completed, the trained classifier was applied to a test sample of datasets. Worry, hate, love, grief, happiness, neutral, anger, boredom, surprise, relief, enthusiasm, fun, and empty are some of the finer emotions expressed in new tweets.

III. COMPARATIVE ANALYSIS

A. Support Vector Machine

The Support Vector Machine, or SVM, is a useful the Supervised Learning technique that possibly used to solve both classification and regression issues. However, for most part it is utilized in Machine Learning for Classification difficulties.
The SVM algorithm's purpose is to discover optimum line or decision boundary for categorising n-dimensional spaces into classes so that the extra data points can be readily placed in the right category in the future. A hyperplane is the name for the optimal choice boundary.
The maximum points/vectors that assist create the hyperplane are chosen via SVM. Support vectors are the utmost instances, and the algorithm is called a Support Vector Machine.

- *There are Two Types of SVM*

Linear SVM: Linear SVM is a classifier that can be used for linearly separable data, which hints that if a dataset is to classified in two main classes using a single straight line, it is called linearly separable data, and the classifier is named Linear SVM.
Non-linear SVM: Non-linear SVM can be used for non-linearly separate data, which hints that if a dataset can't be classified using a straight line, it's non-linear data, and the classifier employed is called Non-linear SVM.

- *SVM Accuracy*

SVM has a prediction accuracy of 99.996 percent, according to [13].

B. Naive Bayes

The simplest and most widely used classifier is the Naive Bayes classifier [14]. The back probability of a category is processed using a Guileless Bayes characterization model based on the report's wording. It is based on a very basic depiction of the archive as a bag of words. Spam filtering, content characterization, assumption, examination, & recommendatory frameworks are just a few of the applications where the Naive Bayes classifier comes in handy. It forecasts unknown categories using the Bayes hypothesis of likelihood. The Bayes Theorem is used to predict the likelihood that a particular set of capabilities will have a place with a specified name. Bigrams from the Twitter data are used as highlights on Naive Bayes for Twitter opinion analysis. It categorizes tweets as negative and positive.

- *Training and Testing the Dataset*

Training datasets entails creating a model and then training or fitting it to the parameters. Testing datasets is done to ensure that they are of good quality and perform well. Splitting a tweet into words is used to analyze it. Algorithms or libraries determine the intensity of words. +ve and -ve words are differentiated based on their intensity. The tweet is +ve if the intensity of +ve phrases is high. It's possible that the tweet will be neutral at times. In that instance, the tweet has no good or bad connotations. As a result, any tweet can be subjected to the necessary conviction analysis.

- *Naive Bayes Accuracy*

Naïve bayes has a prediction accuracy of 90.436 percent, according to [13].

Table 1: Accuracy reviews

Author and year	Dataset	Algorithm	Accuracy
Apoova Agrwal	11.9 manually extracted tweets	Unigram Senti-features Kernel Unigram + Senti-features Kernel + Senti-features	71.4% 71.30% 73.93% 75.39% 74.61%
Seyed-Ali Bahraini	Twitter data on mobile	Unigram feature, support vector machine, naive base, maxent Hybrid Approach	89.77%
Neethu M.S [17]	Twitter data about electronic results	NB support vector machine min entropy ensembled	89% 90% 90% 90%
Dhiraj Gurkhe [18]	Twits ad data	Unigram model Bi gram model Uni + bigram	81% 15% 65.5%
Geetika Gautam [19]	Feedback Twitter dataset	NB Max Entropy support vector machine Semantic analysis	88% 83.8% 85% 89%

IV. CONCLUSION

Twitter is the biggest platform with the source of structured datasets which helped us to do the sentimental analysis. In this review paper I would like to conclude that sentiment analysis is very useful for understanding the positive and negative thoughts of people using the above-mentioned techniques and methodology. This paper also helps us to understand people's psychology on the basis of their tweets and responses. SVM gives the more accuracy results then naïve bayes. If you want to analyze a tweet you should know all its attributes and phases to perform analysis on it for the accurate results.

REFERENCES

[1] Prerna Mishra, Dr. Ranjana Rajnish, Dr. Pankaj Kumar, “Sentiment Analysis of Twitter Data: Case study on Digital India”, InCITe-2016

[2] Rasika Wagh, Payal Punde, “Survey on Sentiment Analysis using Twitter Datasets”, ICECA-2018

[3] Shikha Tiwari, Anshika Verma, Peeyush Garg, Deepika Bansal, “Social Media Sentiment Analysis on Twitter Datasets”, ICACCS-2020

[4] Chirag Kariya, Preeti Khodke “Twitter Sentiment Analysis”, PRMCEAM, Bandera,India, INCET-2020

[5] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt “Sentiment Analysis of Top Colleges in India Using Twitter Data”, ICCTICT-2016

[6] Sahar A. El Rahman, " Sentiment Analysis of Twitter Data", Computer and Information sciences College Princess Nourah Bint Abdulrahman University,

[7] (Mtech): Department of Computer Science and IT. Dr. BAMU Aurangabad, India, ICECA 2018

[8] Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government." ICT Express (2017)

[9] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016)

[10] Kharche, S. R., and Lokesh Bijole. "Review on Sentiment Analysis of Twitter Data." International Journal of Computer Science and Applications 8 (2015)

[11] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015)

[12] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of Twitter data using machine learning approaches and semantic analysis." Contemporary computing (IC3), 2014 seventh international conference on. IEEE, 2014.

[13] Anurag P. Jain, "Sentiments Analysis Of Twitter Dat Using Data Mining", Dept. of Information Technology Pimpri Chinchwad College of Engineering Pune, India,2015 ICIP.

[14] Huma Parveen & Prof. Shikha Pandey "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm". Dept. of Computer Science and Engineering Rungta College of Engineering and Technology Bhilai. India, 2016.

[15] Agarwal, Apoorv. "Teaching the Basics of NLP and ML in an Introductory Course to Information Science." Proceedings of the Fourth Workshop on Teaching NLP and CL. 2013.A

[16] Seyed-Ali Bahrainian and Andreas Dengel, “Sentiment Analysis and Summarization of Twitter Data", Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on.3-5 Dec. 2013.

[17] Neethu, M. & Rajasree, R.. (2013). Sentiment analysis in Twitter using machine learning techniques. 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013. 1-5. 10.1109/ICCCNT.2013.6726818.

[18] Gurkhe, Dhiraj & Pal, Niraj & Bhatia, Rishit. (2014). Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification. International Journal of Computer Applications. 99. 1-4. 10.5120/17430-8274.

[19] Gautam, Geetika & Yadav, Divakar. (2014). Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. 2014 7th International Conference on Contemporary Computing, IC3 2014. 10.1109/IC3.2014.6897213.