The method of Lexicons applied to Sentiment Analysis on Twitter Data to address NLP issues

M.Raghavi . Nagashree N

School of Computer Science and Engineering
REVA University, Bengaluru, India.

Received: 29 August 2022 / Revised: 29 September 2022 / Accepted: 17 October 2022

Abstract – As the world grows, social networks are one of the biggest sources of information where a lot of people interact and communicate. Among the various social media platforms Twitter is one type of social media that is often used. Users tweet their thoughts to the public. Sharing and taking of reviews has been a helpful way to learn about opinions about things. These opinions can lead to Sentiment Analysis, in Twitter there are tweets that can be sentiments. It can be defined as policy, logic, etc. Firstly, a natural language processing-based pre-processes data framework is created to filter tweets. Tokenization sentiment is a technique to be used as a stemming technique. Overall, the process is intended to ascertain what a person thought about a specific tweet they posted. A bag of Words is incorporated to frame a model concept to analyse sentiment. Using this technique, positive and negative tweets can be classified, and also lexicons are used for developing sentiment analysis of tweets.

Index Terms – Tweeter, sentiment analysis, twitter data analysis, NLP

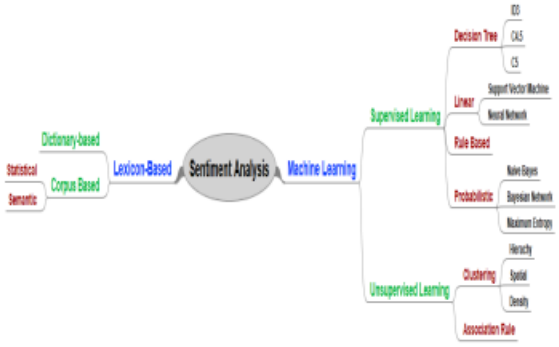
I. INTRODUCTION

Social networking sites are being used for microblogging where it has become a strong tool among Internet users for communication. The popular social networking sites like Twitter, Facebook, Instagram, YouTube etc. are in trend these days. Using Sentiment Analysis, the purpose of the present study is to determine a person's behaviour, mood, and opinion from text data. This stack of text data on Twitter is quite valuable because it stores valuable information. In general, people want opinions from other people to help them make decisions.[1] Finding contradictions between an individual's different emotions is possible with this method. Tweets are limited to 140 characters, which is why users can express their ideas in a short message. By defining positive and negative polarity, it is possible to identify the sentiment of the tweet based on the tweets. Sentiment analysis is carried out using tokenization, stop words removal, and stemming. Later, sentiment is determined using integer multiplication rules and lexicon matching [2].

II. LITERATURE SURVEY

From the researches in the survey, sentiment analysis is not a complicated process, few techniques and procedures should be followed in order to go through with

the analysis [3]. Natural Language Processing



(NLP) is a branch of the field of artificial intelligence. Through natural language, this field provides a link to human interaction. This research presents a method for sentiment analysis from Twitter-on-Twitter data. By using NLP pipeline, we have cleaned redundant information and stop words are removed from tweets for sentiment analysis which has already been discussed in the previous section. This requires more cognitive processing skills. There is an emoticon in a tweet since we do not analyse tweets by emoticons, so the NLP pipeline removes this emoticon and many redundant words from this tweet.

Set of texts represented in NLP is called the corpus. After the initial processing phase, the process continues [4]. The tweet expresses a negative statement about the tweet. In this way, the proposed technique can also be used to give a sentiment for comparative sentences.

In text mining it is known as opinion or sentiment analysis and collecting tweets with that keyword and conducting sentiment analysis on the tweet is very valuable too. Tweets can be structured, semi-structured, and unstructured. The initial processing of data is nothing but filtering data to erase incomplete and noisy data. The most important sentiment indicator is sentiment words [5]. These are words that are commonly used to express positive or negative sentiments. Deleting retweets, Removing URLs, special characters, Punctuation, Numbers, etc., Removing Stop words, Stemming, are some procedures used for sentiment analysis. The levels of sentiment analysis[5] is shown in fig 1.

Fig 1: Levels of sentiment analysis

There are different levels of sentiment analysis.
Document Level Analysis
Sentence Level Analysis

Sentence level analysis in which emotions are analysed at each sentence
Data: - Collecting datasets (tweets) that are done by users.

Pre-processing: - With the aim of improving the performance of analysis, it is necessary to do pre-processing of data before exploring it. Features- For making the sentiment analysis model, we have to extract every single feature from the text data which are broadly categorized into morphological features, word N-gram features, etc. As a classifier in our Sentiment Analysis experiment model, we use support vector machines and train them over the training samples[5]. Above are the methods of data processing and analysis production.

III. METHODOLOGY

After going through different papers on the said topic an accumulative methodology shall be framed. Based on three important components, the method includes three steps: Data Extraction from a specific project or product, pre-processing the extracted Tweets using Natural Language Toolkit (NLTK), and classifying the Tweets using a Classifier model.

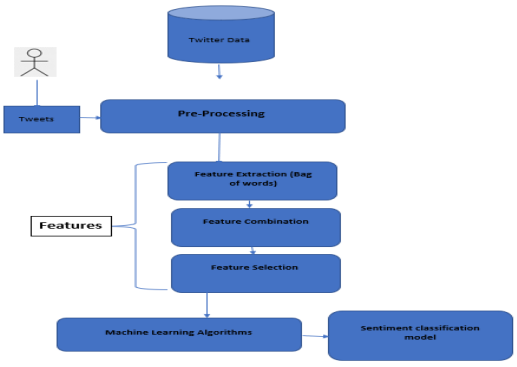
Experimental research is research that manipulates or controls natural situations by creating artificial conditions. In conducting experimental research, there are three aspects

that must be considered, namely control, manipulation, and observation. The Twitter data is using data extraction, tokenization, stemming, lemmatization, stop words removal, parts of speech tagging, named entity recognition, creating a data frame, text modelling as modules for analysis [6]. A natural language processing (NLP) based pre-processed data framework to filter tweets and perform sentiment analysis is developed. The BoW model is used to classify and model text and is fed into the method to analyse tweets. First it is needed for a classifier model that can differentiate tweets:

Suggested algorithm for a classifier model.
Step 1: Importing dataset
Step 2: Serialization and de-serialization
Step 3: save to a file and then load again
Step 4: FOR each data in the dataset:
Step 5: extract from and append required data into dataset
Step 6: Create a simple binary bag of words model
Step 7: Create a simple TF-IDF model
Step 8: Create training and testing datasets:
Step 9: Create a classifier model using logistic regression
Step 10: apply classifier model to analyse
Step 11: vectorizer ← TFIDFVectorizer ()
Step 12: Unpick the classifier and vectorizer for read operation

The proposed framework for sentiment analysis using Lexicons is depicted in fig 2.

Fig 2: Architecture of Sentimental analysis using Lexicons



IV. RESULTS AND DISCUSSIONS

The data is stored in the database in the form of text. As a result of the analysis, a list of adjectives is developed along with manual labelling that is used as a lexicon and for sentiment analysis. The results of this design are then used as a basis to build sentiment analysis in the construction phase of the system. A user initially inputs the topic so that there are some tweets that match the topic and the topic will be parsed through the Twitter API to the server so that some tweets that contain words according to the topic are obtained [7,8].

The next step will be tokenizing and slang removal: This stage attempts to replace words that have current terms with standard words according to language rules. If it matches, it will be replaced with If the word does not match, the word is omitted. After this the stemming is done in order get the basic word is obtained using Porter Stemming which is in the Python library. With the above features we can evaluate the score of the sentences and tweets as -1 and 1 which will show whether the sentiment behind is positive or negative. For example: "Her fashion sense is not good". Here the score for "not* good" is -1, so it's a negative sentence. The data distribution for negative and positive classifiers are shown in fig 3 as below.

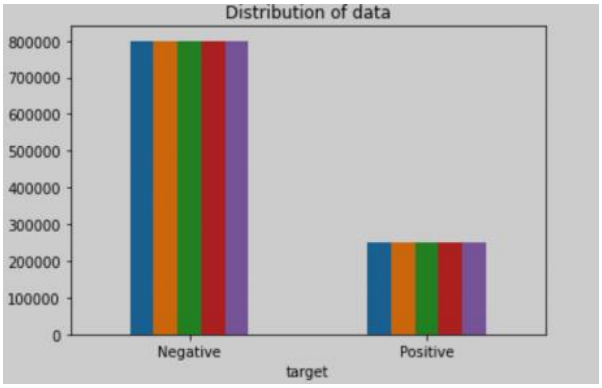


Fig 3: Data distribution graph for negative and positive target sets in Lexicons

V. CONCLUSION

It is understandable the importance of the opinion of people for the product-based company or public. A fine-grained sentiment analysis of social media sites like Twitter is set. The classifier built can be utilized as a data analysis tool in NLTK. A Lexicon approach shall be used for further Sentiment Analysis to ensure accuracy. The evaluation of feature extraction from the Twitter data is successfully completed and can be applied for further analysis and decision making. These methods can be used along with SVMs to increase the accuracy of the model in the future.

REFERENCES

1. Tiwari, S., Verma, A., Garg, P., & Bansal, D. (2020, March). Social media sentiment analysis on Twitter datasets. In *2020 6th international conference on advanced computing and communication systems (ICACCS)* (pp. 925-927). IEEE.
2. Rosenthal, S., Farra, N., & Nakov, P. (2019). SemEval-2017 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.00741*.
3. Ahmed, S. S. T., Thanuja, K., Guptha, N. S., & Narasimha, S. (2016, January). Telemedicine approach for remote patient monitoring system using smart phones with an economical hardware kit. In *2016 international conference on computing technologies and intelligent data engineering (ICCTIDE'16)* (pp. 1-4). IEEE.
4. Ahmed, K., El Tazi, N., & Hossny, A. H. (2015, October). Sentiment analysis over social networks: an overview. In *2015 IEEE international conference on systems, man, and cybernetics* (pp. 2174-2179). IEEE.
5. Nagashree, N., Patil, P., Patil, S., & Kokatanur, M. (2019). Performance metrics for segmentation algorithms in brain MRI for early detection of autism. *Int. J. Innovative Technol. Exploring Eng.(IJITEE)*, *9*.
6. Nagesh, N., Patil, P., Patil, S., & Kokatanur, M. (2022). An architectural framework for automatic detection of autism using deep convolution networks and genetic algorithm. *International Journal of Electrical & Computer Engineering (2088-8708)*, *12*(2).
7. Ahmed, S. T., Singh, D. K., Basha, S. M., Abouel Nasr, E., Kamrani, A. K., & Aboudaif, M. K. (2021). Neural network based mental depression identification and sentiments classification technique from speech signals: A COVID-19 Focused Pandemic Study. *Frontiers in public health*, 1926.
8. Sreedhar Kumar, S., Ahmed, S. T., & NishaBhai, V. B. Type of Supervised Text Classification System for Unstructured Text Comments using Probability Theory Technique. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(10).
9. Al-Shammari, N. K., Alzamil, A. A., Albadarn, M., Ahmed, S. A., Syed, M. B., Alshammari, A. S., & Gabr, A. M. (2021). Cardiac Stroke Prediction Framework using Hybrid Optimization Algorithm under DNN. *Engineering, Technology & Applied Science Research*, *11*(4), 7436-7441.