ISSN 2395-1621

A Survey on Sentiment Analysis for Tweets using Patterns and Strategies to Detect the Fake Tweets

[#1]Chavan Prajakta B., [#2]Godase Snehal R., [#3]Kadam Shivkanya K., [#4]Mohite Pranita D.

prajaktachavan017@gmail.com
snehalgodase99@gmail.com
shivkanyakadam67@gmail.com
pranitamohite35@gmail.com

Department of Information Technology
Vidya Pratishthans Kamalnayan Bajaj Institude Of Engineering & Technology, Baramati.

ABSTRACT

Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd. Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. Knowledge base approach and Machine learning approach are the two strategies used for analyzing sentiments from the text. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. Twitter offers organizations a fast and effective way to analyze customers' perspectives toward the critical to success in the market place. Developing a program for sentiment analysis is an approach to be used to computationally measure customers' perceptions. This project uses knowledge base including various patterns for tweets along with multiple strategies to detect the sentiment expressed in a tweet and if a tweet is genuine or not. Various machine learning and knowledge base approaches are used to compare patterns and apply strategies and NLP for sentiment analysis.

Keywords: NLP (Natural Language Processing), Sentiment analysis, machine learning, influence of tweets, POS(Part of Speech).

ARTICLE INFO

Article History

Received: 28th April 2021

Received in revised form : 28th April 2021

Accepted: 1st May 2021

Published online : 2nd May 2021

I. INTRODUCTION

Twitter has emerged as a major micro-blogging website, having over 100 million users generating over 500 million tweets every day. With such large audience, Twitter has consistently attracted users to convey their opinions and perspective about any issue, brand, company or any other topic of interest. Due to this reason, Twitter is used as an informative source by many organizations, institutions and companies. On Twitter, users are allowed to share their opinions in the form of tweets, using only 140 characters. This leads to people compacting their statements by using slang, abbreviations, emoticons, short forms etc. Along with this, people convey their opinions by using sarcasm and polysemy. Hence it is justified to term the Twitter language as unstructured. In order to extract sentiment from tweets, sentiment analysis is used. The results from this can be used in many areas like analyzing and monitoring changes of sentiment with an event, sentiments regarding a particular brand or release of a particular product, analyzing public view of government policies etc.

A lot of research has been done on Twitter data in order to classify the tweets and analyze the results. In this project we aim to predict the sentiments from tweets by checking the polarity of tweets as positive, negative or irrelevant. Sentiment analysis is a process of deriving sentiment of a particular statement or sentence. It's a classification technique which derives opinion from the tweets and formulates a sentiment and on the basis of which, sentiment classification is performed. Sentiments are subjective to the topic of interest. We are required to formulate that what kind of features will decide for the sentiment it embodies. In the programming model, sentiment we refer to, is class of

entities that the person performing sentiment analysis wants to find in the tweets. The dimension of the sentiment class is crucial factor in deciding the efficiency of the model. For example, we can have two-class tweet sentiment classification (positive and negative) or three class tweet sentiment classification (positive, negative and irrelevant). Sentiment analysis approaches can be broadly categorized in two classes – lexicon based and machine learning based. Lexicon based approach is unsupervised as it proposes to perform analysis using lexicons and a scoring method to evaluate opinions. Whereas machine learning approach involves use of feature extraction and training the model using feature set and some dataset.

The basic steps for performing sentiment analysis includes data collection, pre-processing of data, feature extraction, selecting baseline features, sentiment detection and performing classification either using simple computation or else machine learning approaches.

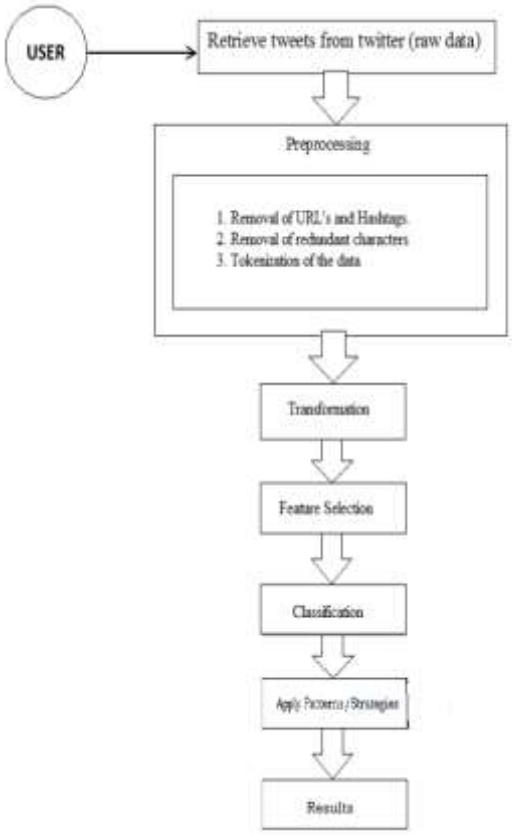

Fig 1. System architecture diagram

In this system will retrieve tweets from twitter using twitter API based on the query. The collected tweets will be subjected to preprocessing. We will then apply the various patterns and strategic algorithms including some of machine learning algorithms for NLP to supervised the data. The results of the algorithms i.e. the sentiment and influence will be represented in graphical manner (pie charts/bar charts). The proposed system is more effective than the existing one. This is because we will be able to know how the statistics determined from the representation of the result can have an impact in a particular field as well as influence of negativity spread by rumors.

II. REVIEW OF LITERATURE

Parikh and Movassate [1] implemented two Naive Bayes unigram models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets. They found that the Naive Bayes classifiers worked much better than the Maximum Entropy model could.

Go et al. [2] proposed a solution by using distant supervision, in which their training data consisted of tweets with emoticons. This approach was initially introduced by Read [3]. The emoticons served as noisy labels. They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM). Their feature space consisted of unigrams, bigrams and POS. The reported that SVM outperformed other models and that unigram were more effective as features.

Pak and Paroubek [4] have done similar work but classify the tweets as objective, positive and negative. In order to collect a corpus of objective posts, they retrieved text messages from Twitter accounts of popular newspapers and magazine, such as “New York Times”, “Washington Posts” etc. Their classifier is based on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features.

Barbosa et al. [5] too classified tweets as objective or subjective and then the subjective tweets were classified as positive or negative. The feature space used included features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words.

Bifet and Frank [6] used Twitter streaming data provided by Firehouse, which gave all messages from every user in real-time. They experimented with three fast incremental methods that were well-suited to deal with data streams: multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. They concluded that SGD-based model, used with an appropriate learning rate was the best.

III. CONCLUSION

The system set out to solve a practical problem of sentiment analysis and genuinely check of Twitter posts. This system proposed a method using knowledge base patterns, strategies and machine learning approaches. These methods are proposed to increase the accuracy of sentiment check for tweets. Patterns can be used to evaluate if the tweets was a influenced rumor or a genuine post by any user. By using API of twitter it is possible to work on live tweets than to work on offline data. Querying and fetching of particular tweets from twitter is possible by using its API. Finding influence or negativity spread by users can be useful in various analytical tasks.

IV. 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 UserGenerated Twitter Updates using Various Classification Techniques", CS224N Final Report, 2009

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

[4] J. Read. "Using emoticons to reduce dependency in machine learning techniques for sentiment classification".In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005

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

[6] A. 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.

