iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 7 **Issue:** VI **Month of publication:** June 2019

**DOI:** http://doi.org/10.22214/ijraset.2019.6003

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177*
*Volume 7 Issue VI, Jun2019- Available at www.ijraset.com*

# Sentiment Anaysis in Twitter

Miss. Hema R. Khemnar[1], Miss. Samiksha S. Unawane[2], Miss. Priya S. Shelke[3], Mr. S. K. Korde[4]
[1, 2, 3, 4]*Department of Computer Engineering Pravara Rural Engineering College Loni, Savitribai Phule Pune university*

***Abstract:*** ***With the rapid growth of social networks and microblogging websites, communication between people from different cultural and psychological backgrounds became more direct, resulting in more and more "cyber" conflicts between these people. Consequently, hate speech is used more and more, to the point where it became a serious problem invading these open spaces.***
***Hate speech refers to the use of positive or negative language, targeting a specific group of people sharing a common property, whether this property is their gender, their ethnic group or race or their believes and religion, etc. While most of the microblogging websites and online social networks forbid the use of hate speech, the size of these networks and websites makes it almost impossible to control all of their content. Therefore, it is necessary to detect such speech automatically and filter any content that presents hateful language. In this project we propose an approach to detect sentiment analysis in Twitter. Our approach is based on positive and negative words that are automatically collected from the training set. Our experiments show that our approach detecting whether a tweet is positive or negative (binary classification).***
***Keywords: Tweet, Segmentation, machine learning, Sentiment analysis, Sentiment classification.***

## I. INTRODUCTION

### *A. Motivation*

Apache spark is the currently used by many big companies to access data fast. It makes use of big data to access the data from various clusters of Twitter Server and saves time. So we thought of using this technology in our project to access data from twitter server to access tweets.

### *B. Problem Statement*

Today Social Networking Sites (SOCIAL NETWORKING SITES) have become an important part of our day to day life. We share a lot of personal data on these sites. They help us to make the world smaller and integrate like a small village with each other. There are many SOCIAL NETWORKING SITES available today and many more are increasing each day. Thus a user uses many Social Networking Sites each day and communicate and share data with friends and family. This communication medium gave rise to complex structure whether a user really like the Social Networking Sites which he uses more or he needs another Social Networking Sites other than he uses more. Thus one of the most famous SOCIAL NETWORKING SITES is TWITTER which is used to share data and post our thoughts and latest buzz upon the internet. The users using TWITTER have increased constantly in the recent years. So the analysis of this SOCIAL NETWORKING SITES may help in answering and predicting many answers.

## II. LITERATURE SURVEY

This section summarizes some of the scholarly and research works in the field of Machine Learning and data mining to analyze sentiments on the Twitter and preparing prediction model for various applications. As the available social platforms are shooting up, the information is becoming vast and can be extracted to turn into business objectives, social campaigns, marketing and other promotional strategies as explained in . The benefit of social media to know public opinions and extract their emotions and explained how twitter gives advantage politically during elections. Potato Leaf Diseases Detection and Classification System The concept of hashtag is used for text classification. They suggested how previous research work suffered from lack of training set and misses some features of target data. They opted two stage approach for their framework- first preparing training data from twitter using mining conveying relevant features and then propounding the Supervised Learning Model to predict the results of Elections held in USA in 2016. After collecting and preprocessing the tweets, training data set was created first by manual labelling of hashtags and forming clusters, next by using online Sentimental Analyzer VADER which outputs the polarity in percentage. This approach reduced the number of tweets or training set and further they applied Support Vector Machine and Naive Bayes classification algorithm to determine the polarity of tweets. Multistage Classification approach was used where an entity classifier receives general class of tweets and categorise them with respect to individual candidates for comparison. The metric they used to determine the winner was the "PvT ratio" which is Positive number of tweets to total count of tweets for respective candidate.

Sentiment Analysis by researchers Imran et al. exploited the technology 'Apache Spark' for fast streaming of tweets and presented the approach StreamSensing to handle real time data in unstructured and noisy form. They conducted the approach on twitter data to find some useful and interesting trends which further can be generalized to any real-time text stream. Unsupervised learning approach is used to locate interesting patterns and trends from tweets processed on Apache Spark. Inspired by the approach described by Zhu et al. and Li et al. for mining data by selecting time window, authors opted for sliding window method for capturing the live streams of tweets. The common approach found in almost all relevant research works constitutes data collection using Twitter API, preprocessing of data, filtering of data then approaches in feature extraction, classification and pattern analysis makes the distinction.

## III.PROPOSED SYSTEM

*A. Goal And Objectives*

*1)* To make effective API
*2)* To make effective Apache spark
*3)* Remove stop words.
*4)* Apply Machine Learing

*B. Statement Of Scope*

*1) Sentiment Analysis:* The System can be used to analyze the differentiate between a useful and non-useful tweets.
*2) Product Ranking:* The System can be used to analyze the rating of products and differentiate between a useful and non-useful tweets.

*C. System Architecture*

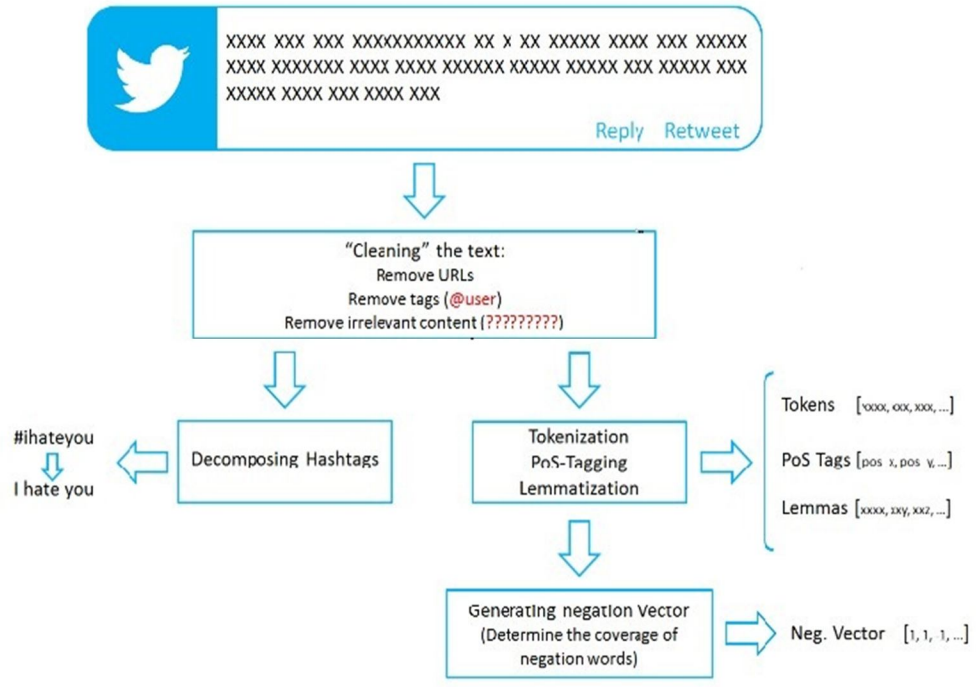



Fig: System Architecture

In a first step, we clean up the tweets. This includes the removal of URLs (which starting either with "http://" or "https://") and tags (i.e., "@user") and irrelevant expressions (words written in languages that is not supported by ANSI coding). In Lemmatization used both token POS tags. we can use the NLP(Natural Language Processing) task of tokenization and lemmatization. However to perform the POS tagging with rely on gate twitter POS tagger. This is because OpenNLP presents poor performances on PoS tagging of informal and noisy texts such as tweets.

*D. Algorithm*

*1) SVM (Support Vector Machine):* In machine learning, support vector machines (SVMs, also support vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.When data is unlabelled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The support vector clustering algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications.

## IV. APPLICATIONS

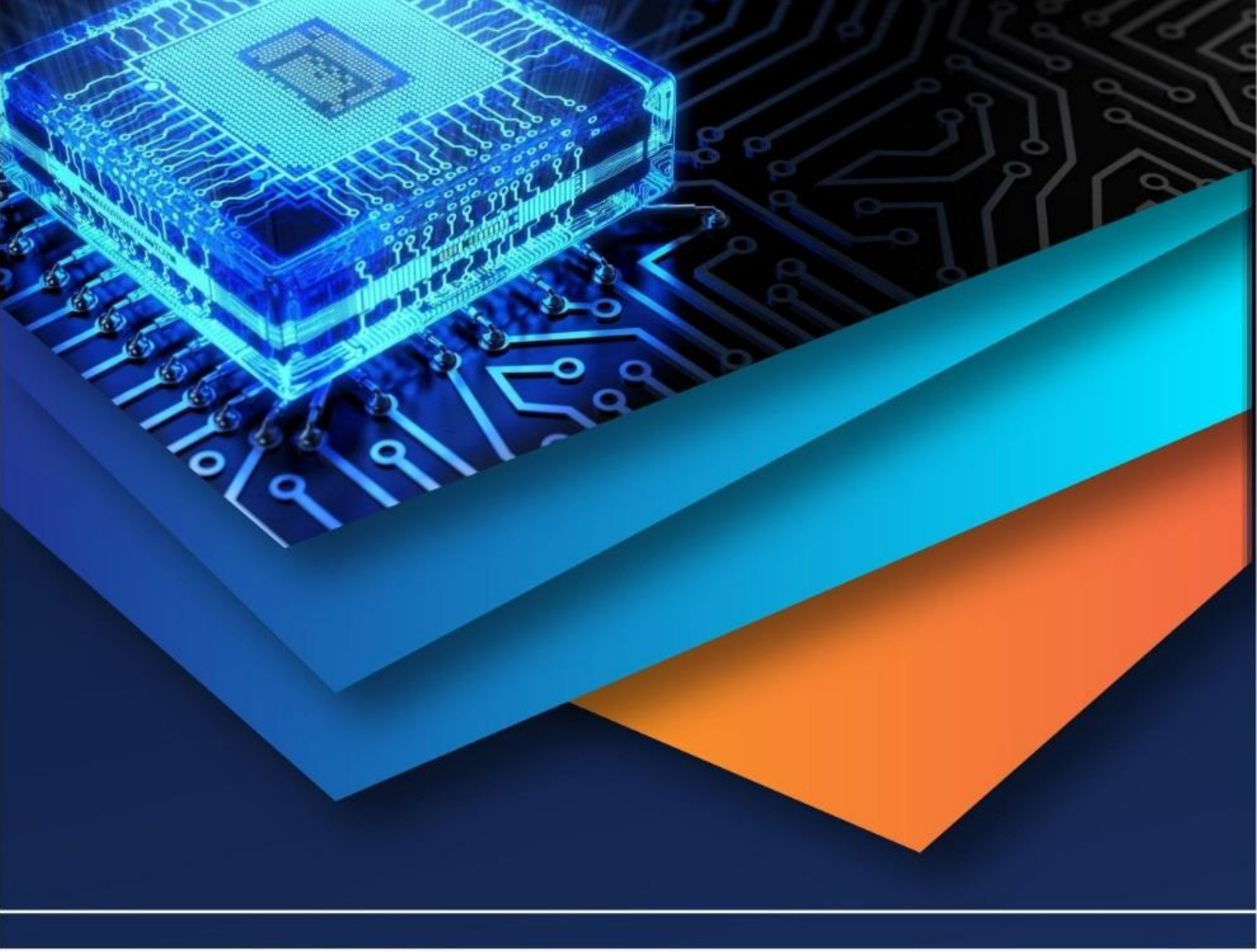
*A.* Review related websites.
*B.* Sub component technology.
*C.* Business and government intelligence.
.

## V. CONCLUSION

In this project, we are developing novel sentiment analysis approach using TWEETER and Apache Spark together. The basic idea of the project is to use distributed computing in training and testing the machine learning classification using named nodes and data nodes together. We are going to assemble various predictions by machine learning algorithms together and view the results in two classes such as positive, negative according to the predictions returned by the system.

## REFERENCES

[1] Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection Hajime Watanabe, Mondher Bouazizi, Tomoaki Ohtsuki
[2] Peter J. Breckheimer, “A Haven for Hate: The Foreign and Domestic Implications of Protecting Internet Hate speech Under the First Amendment,” in South California Law Review, vol. 75, no. 6, Sep. 2002.
[3] P. Burnap, and M. L. Williams, “Cyber Hate Speech on Twitter: An Application of Machine Classification and Statisitcal Modeling for Policy and Decistion Making,” in Policy and Internet pp. 223–242, June 2015.
[4] A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin, “Offensive Language Detection Using Multi-level Classification,” Advances in Artificial Intelligence, vol. 6085, pp. 16–27, Springer, Ottawa, Canada, June 2010.
[5] D. King and G.M. Sutton, “High Times for Hate Crime: Explaining the Temporal Clustering of Hate Motivated Offending”, in Criminology pp. 871–894, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY