International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Computational Linguistic Identification and Opinion Mining On Twitter (CLIOMOT)

Prathmesh Gavhalkar[1], Ammarah Khan[2], Rasika Halkude[3], Jyoti Ghodekar[4], Vishnu Kamble[5], Digvijay Patil[6]

[1,2,3,4] *Student, Dept. of Information Technology Engineering, P.E.S's Modern College of Engineering, Pune, Maharashtra, India*
[5,6] *Asst. Professor, Dept. of Information Technology Engineering, P.E.S's Modern College of Engineering, Pune, Maharashtra, India*

Abstract - *Twitter, as a social media is a very popular way of expressing opinions and interacting with other people in the online world. When taken in aggregation tweets can provide a reflection of public sentiment towards events. In this paper, we provide a positive or negative sentiment on Twitter posts using a well-known machine learning method for text categorization. In addition, we use manually labeled (positive/negative) tweets to build a trained method to accomplish a task. The task is looking for a correlation between twitter sentiment and events that have occurred. The trained model is based on the Naive Bayes and Support Vector Machi(SVM) classification method.*

Also we used external lexicons to detect subjective or objective tweets, added Unigram and Bigram features and used TF-IDF (Term Frequency-Inverse Document Frequency) to filter out the features. we used Twitter Streaming API and some of the official hash tags to mine, filter and process tweets, in order to analyze the reflection of public sentiment towards unexpected events. The same approach, can be used as a basis for predicting future events. In the context of a twitter sentiment analysis, at its simplest, sentiment analysis quantifies the mood of a tweet or comment by counting the number of positive and negative words.

***Keywords*: Twitter Streaming API, Opinion Mining, NLP, Sentiment Analysis, Naive Bayes, SVM, BLR.**

1. INTRODUCTION

Twitter, one of the most common online social media and micro-blogging ser-vices, is a very popular method for expressing opinions and interacting with other people in the online world. Twitter messages provide real raw data in the format of short texts that express opinions, ideas and events captured in the moment. Tweets (Twitter posts) are well-suited sources of streaming data for opinion mining and sentiment polarity detection. Opinions, evaluations, emotions and speculations often reflect the states of individuals; they consist of opinionated data expressed in a language composed of subjective expressions.

Social media is emerging rapidly on the internet. This media knowledge helps people, company and organizations to analyze information for important decision making. Opinion mining is also called as sentiment analysis which involves in building a system to gather and examine opinions about the product made in reviews or tweets, comments, blog posts on the web. Sentiment is classified automatically for important applications such as opinion mining and summarization.

To make valuable decisions in marketing analysis where implement sentiment classification efficiently. Reviews contain sentiment which is expressed in a different way in different domains and it is costly to annotate data for each new domain. The analysis of online customer reviews in which firms cannot discover what exactly people liked and did not like in document-level and sentence-level opinion mining. So, now opinion mining ongoing research is in phrase-level opinion mining. It performs finer- grained analysis and directly looks at the opinion in online reviews. The proposed system is based on phrase-level to examine customer reviews. Phrase-level opinion mining is also well-known as aspect based opinion mining. It is used to extract most important aspects of an item and to predict the orientation of each aspect from the item reviews. The projected system implements aspect extraction using frequent item set mining in customer product reviews and mining opinions whether it is positive or negative opinion. It identifies sentiment orientation of each aspect by supervised learning algorithms in customer reviews.

Data mining research has successfully shaped numerous methods, tools, and algorithms for handling huge volume of data to solve real world problems. The key objectives of the data mining process are to effectively handle large-scale data, mine actionable rules, patterns and gain insightful knowledge. The explosion of social media has created extraordinary opportunities for citizens to publicly voice their opinions. Because social media is widely used for diverse purposes, huge content of user created data exist and can be made an accessible for data mining. Recent researches in data mining focus on opining mining.

1.1 Literature Review

In the past decade, new forms of communication, such as micro blogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them.

In the existing system the algorithms and the techniques used were complex and difficult to understand and results were not much accurate and consistent. The proposed system consists of methods and techniques which have accurate and consistent results and also are easy to understand. Complexity in the solving the problems is reduced due to consideration of the proposed system.

Tweets and texts are short: a sentence or a headline rather than a document. The language used is very informal, with creative spelling and punctuation, misspellings, slang, new words, URLs, and genre-specific terminology and abbreviations, such as, RT for "re-tweet" and hash tags, which are a type of tagging for Twitter messages.

Table -1:

Paper Title	Part Used
Sentiment Analysis on Twitter using Streaming API	NLTK, Twitter API, Different Phases
Sentiment Analysis of Tweets using Machine Learning Approach	sentiment analysis on social media
Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm	Framework Implementation of Naïve Bayes Algorithm
Comparative Analysis of Sentiment Orientation Using SVM and Naive Bayes Techniques	Methodology: Dataset, Text Processing, Porter Algorithm.
Comparative Study of Classification Algorithms used in Sentiment Analysis	Classifications: Naïve Bayes, Max Entropy, Boosted Trees, Random Fores

2. DESIGN AND ARCHITECTURE

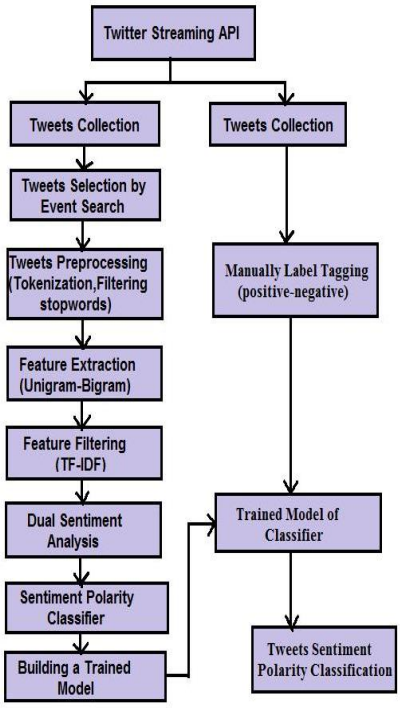


Figure 2: Architecture of the system

2.1 Tweets Collection:

The first module comprises of Data gathering for the creation of a training set and then collecting tweets from a particular event. Data gathering is made up of two steps using Twitter streaming API: The first is collecting the data to use as a training set to building the model. This consisted of number of Tweets manually labelled "Positive" or "Negative". The second steps are collecting Tweets during any particular event and classify them according to some of the official Hash tags. In addition, the twitter usernames of concerned personalities related to that event are used to extract tweets relating to events. The data is in JSON format as a set of documents.

2.2 Tweets Text Pre-processing:

As a first step towards finding a tweets sentiment and in order to obtain accurate sentiment classification, we needed to filter out noise and meaningless symbols that do not contribute to a tweets sentiment from the original text.

2.3 Feature Extraction:

Selecting a useful list of words as features of a text and removing large number of words that do not contribute to the texts sentiment is defined as feature extraction.

2.4 Feature Filtering:

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistical method to filter the features by

weighting and scoring each of the unigrams and N-grams using the frequency of words in the text.

2.5 Dual Sentiment Analysis:

First, we strengthen the DSA algorithm by adding a selective data expansion procedure. Second, we extend the DSA framework from sentiment polarity classification to positive-negative-neutral sentiment classification. Third, we propose a corpus-based method to construct a pseudo-antonym dictionary that could remove DSAs dependency on an external antonym dictionary.

2.6 Sentiment Classifier:

This method selects 90% for the training set, and 10% for the testing set, repeating it on 10 different sections of dataset.

2.7 Building a Trained model:

Labelling an opinionated text and categorizing it overall into a positive or negative class is called sentiment polarity classification. The neutral label is used for more objective items that have lack of opinion in the text, or where there is a mixture of positive and negative opinions therein. We need to use all the subjective tweets, including positive or negative sentiment. There are methods of extracting the useful words in order to detect the sentiment of tweets.

3. REQUIREMENT ANALYSIS

This section describes about the requirements. It specifies the hardware and software requirements that are required in order to run the application properly.

3.1 System requirement specification:

A structured collection of information that embodies the requirements of a system. A **Business Analyst**, sometimes titled system analyst, is responsible for analysing the business needs of their clients and stakeholders to help identify business problems and propose solutions. Within the systems development life cycle domain, the BA typically performs a liaison function between the business side of an enterprise and the information technology department or external service providers.

3.2 Feasibility:

A system can be developed technically and that will be used if installed must still be a good investment for the organization. In an Economical feasibility, the development cost in creating the system is evaluated against the ultimate benefit derived from the new systems. Financial benefits must equal or exceed the costs. The system is economically feasible. It does not require any addition hardware or software. Since the interface for this system is developed using the existing resources and technologies available at NIC, there is nominal expenditure and Economical feasibility for certain.

3.3 Functional Requirements:

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality.

3.4 Non-Functional Requirements:

Non Functional requirements, as the name suggests, are those requirements that are not directly concerned with the specific functions delivered by the system. They may relate to emergent system properties such as reliability response time and store occupancy. Alternatively, they may define constraints on the system such as the capability of the Input Output devices and the data representations used in system interfaces. Many non-functional requirements relate to the system as whole rather than to individual system features.

3.5 Software Requirement:

[1]Operating System: Windows

[2]Technology: Python(3.6)

[3]IDE: NLTK(JetBrains PyCharm 2020)

3.6 Hardware Requirement:

[1] Hardware: Pentium

[2] Speed: 1.1 GHz

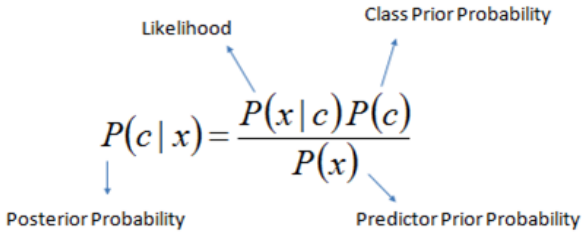
[3] RAM: 1 GB

[4]Hard Disk: 20GB

4. ALGORITHMS OF PROPOSED SYSTEM

4.1 Introduction to Naive Bayes Algorithm:

Naive Bayes is basically a probabilistic approach for the classification of tweets on the basis of their polarity in terms of either a positive opinion or a negative opinion.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Figure 4.1: Naive Bayes Classification

A Multinomial event model is used which counts word occurrences. How-ever, tweets are such short texts that rarely contain multiple occurrences of the same word, thus Bernoulli was adequate for sentiment analysis. A disadvantage would be that conditional independence

often does not hold in reality for text, yet this model performed fairly well for this concept.

In this section, we present the implementation of our Hadoop framework for efficiently executing Naive Bayes algorithm. To implement Naive Bayes algorithm we need a trained SentiWordNet dictionary which is available online. It consists of collection of different word with its synonym and its polarity. The synonym represents the similar word meaning which will be having same polarity. The polarity represents the positivity of the word in the context of the sentence.

4.2 Introduction to SVM Algorithm:

We see that SVM is one of the most accurate classifier. As far as sentiment analysis is concerned an multiclass SVM with a one vs one scheme is considered, which creates one binary classifier for every different pair of classes, and then uses an opinion mechanism to ultimately choose one class. One drawback of SVM is that if the size of the feature vector is larger than the number of training samples, it tends to over fit on the training data, which lowers the accuracy on the testing data. Another drawback of SVM occurs with neutral opinion, which could pose an issue to our project due to the small size of our dataset.

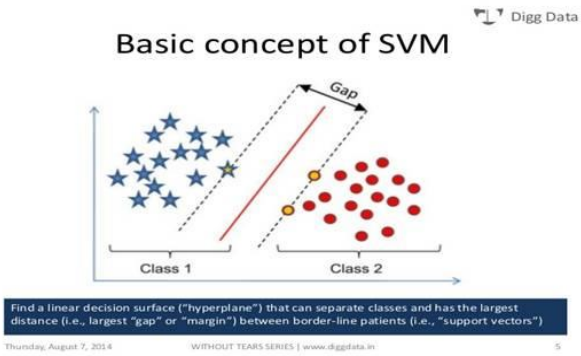


Figure 4.2: Basic Concept of SVM.

4.3 Introduction to BLR Algorithm:

Bayesian Logistic regression provides a positive or negative sentiment on Twitter posts as it is a well-known machine learning method for text categorization. In addition, we use manually labelled (positive/negative) tweets to build a trained method to accomplish a task. The task is looking for a correlation between twitter sentiment and events that have occurred. The trained model is based on the Bayesian Logistic Regression (BLR) classification method.

In statistics the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1 and the sum adding to one.

5. SMILEY DETECTION:

Things which we are adding to the Proposed Technique: Nowadays in most of the tweets people are using smileys in order to express their opinions. So, in order to perform sentiment analysis we should be in a position to classify those tweets on the basis of opinion polarity which consists of various kinds of smileys or emoticons.

Figure 5: Smily Detection.

6. RESULT

This work is of tremendous use to the people and industries which are based on sentiment analysis. For example, Sales Marketing, Product Marketing etc. The key features of this system are the training module which is done with the Classification based on Nave Bayes ,SVM, BLR Time Variant Analytics and the Continuous learning System.

The fact that the analysis is done real time is the major highlight of this paper. Several existing systems store old tweets and perform sentiment analysis on them which gives results on old data and uses up a lot of space. But in this system, the tweets are not stored which is cost effective as no storage space is needed. Also all the analysis is done on tweets real-time. So the user is assured that, getting new and relevant results

7. CONCLUSION

Sentiment polarity measures for various entities and events to see how positively or negatively people react or talk about them. Dual Sentiment Analysis (DSA) model is very effective for polarity classification and it significantly outperforms several alternative methods of considering polarity shift. Creating reversed reviews to assist supervised sentiment classification.

In addition, we strengthen the DSA algorithm by developing a selective data expansion technique that chooses training reviews with higher sentiment degree for data expansion. The experimental results show that using a selected part of training reviews for data expansion can yield better performance than that using all reviews.

REFERENCES

[1] ”Sentiment Analysis on Social Media” IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining-Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas.

[2] ”MapReduce programming with apache Hadoop”M. Bhandarkar-International Symposium on Parallel Distributed Processing (IPDPS).

[3] “Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques.” Shweta Rana, Archana Singh.

[4] . "Review on Developing Corpora for Sentiment Analysis Using Plutchik’s Wheel of Emotions with Fuzzy Logic. “International Journal of Computer Sciences and Engineering (IJCSE) 2 (2013): 14-18. Chafale, Dhanashri, and Amit Pimpalkar.

[5] "Tokenization and Filtering Process in RapidMiner." International Journal of Applied Information Systems (IJAIS)–ISSN (2014): 2249-0868. Verma, Tanu, and Deepti Gaur Renu.

[6] https://www2.bui.hawhamburg.de/pers/ursula.schulz/astep/porter.pdf

[7] "Sentiment analysis and opinion mining: a survey." International Journal 2.6 (2012). Vinodhini, G., and R. M. Chandrasekaran.

[8] “Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorith” Huma Parveen, Prof. Shikha Pandey.

[9] "Sentiment Analysis on Social Media", 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas.

[10] https://en.wikipedia.org/wiki/Sentiment_analysis

[11] https://en.wikipedia.org/wiki/Text_processing

[12] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[13] https://en.wikipedia.org/wiki/Support_vector_machine

BIOGRAPHIES

Prathmesh Gavhalkar
Student at Dept. of Information Technology, P.E.S’s Modern College of Engineering, Pune, Maharashtra, India

Ammarah Khan
Student at Dept. of Information Technology, P.E.S’s Modern College of Engineering, Pune, Maharashtra, India

Rasika Halkude
Student at Dept. of Information Technology, P.E.S’s Modern College of Engineering, Pune, Maharashtra, India

Jyoti Ghodekar
Student at Dept. of Information Technology, P.E.S’s Modern College of Engineering, Pune, Maharashtra, India

Vishnu. S. Kamble
Asst. Professor at Dept. of Information Technology, P.E.S’s Modern College of Engineering, Pune, Maharashtra, India

Digvijay. A. Patil
Asst. Professor at Dept. of Information Technology, P.E.S’s Modern College of Engineering, Pune, Maharashtra, India