Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approaches

Ankit Pradeep Patel
UG Student
Department of Information Technology
K. J. Somaiya Institute Engineering & Information Technology Mumbai, India

Ankit Vithalbhai Patel
UG Student
Department of Information Technology
K. J. Somaiya Institute Engineering & Information Technology Mumbai, India

Sanjaykumar Ghanshyambhai Butani
UG Student
Department of Information Technology
K. J. Somaiya Institute Engineering & Information Technology Mumbai, India

Prashant B. Sawant
Assistant Professor
Department of Information Technology
K. J. Somaiya Institute Engineering & Information Technology Mumbai, India

Abstract

In today's world, micro-blogging sites has become a platform for individuals or organizations across the world to express their opinions, sentiment and experience in the form of tweets, status updates, blog posts, etc. This platform has no political and economic restrictions. This paper discusses an approach where a published stream of tweets on electronic products from the twitter micro-blogging site are then subjected to preprocessing and classified based on their emotional content as positive, negative and neutral. The performance of the unsupervised algorithm is then analyzed. The paper concludes with the comparison of the existing system with the proposed systems and applications of the research.

Keywords: Classification, Data Preprocessing, Machine Learning, Sentiment Analysis

I. INTRODUCTION

Twitter is a micro-blogging website that has become increasingly popular with the network community. Users update short messages, also known as Tweets, which are limited to 140 characters. Users update their personal opinions on many subjects, discuss current topics and write about life events through tweets. This platform is favored by many users because it has no political and economic restrictions and is easily available to large number of people. As the amount of users increase, micro-blogging platforms are becoming a place to find strong viewpoints and sentiment. People use twitter to forecast and analyze in a lot of different areas. For example, people have already forecasted the stock market success by using data from Twitter [7]. People use Twitter to forecast popularity and sales revenue of electronic products. From these case studies, we can know that Twitter is really useful for predicting products, services, or markets. It is one important reason why Twitter is taken into consideration to predict how people think about the popularity of day to day products. Another reason is because Twitter serves as a worthy platform for sentiment analysis due to its large user base with people across the world having different perspective. Twitter contains enormous amount of tweets, with millions being added every day. This can be easily collected through its APIs, which makes it easy to build a training set.

In section II, we did literature survey on this topic which includes the study on the collection of twitter data through the twitter API, pre-processing of the collected data, various unsupervised algorithms and its efficiency. We will consider trending electronic products. Paper[1] has discussed sentiment analysis on the customer's review using classification. In section III, we did research and study on existing system in which we have noticed that the research conducted using the supervised algorithms have some drawbacks. The drawbacks are synonym word vectors and results based on assumptions. These drawbacks have an impact on the efficiency of algorithms. In section IV, we proposed a system which will overcome the problem. In this system we will use the unsupervised algorithms. In section V, we did comparison between existing system and proposed system where one can easily get the newly added features of proposed system and drawbacks of existing system.

II. BACKGROUND WORK

Sentiment analysis deals with identifying and classifying opinions or sentiments which are present in source text. Social media is generating a huge amount of sentiment rich data in the form of tweets, status updates, reviews and blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd. Twitter sentiment analysis is arduous as

All rights reserved by *www.ijirst.org*

compared to basic sentiment analysis due to the presence of slang words and misspellings. The maximum limit of characters that are allowed in Twitter is 140. Machine learning approach can be used for analyzing sentiments from the text.

Some sentiment analysis are performed by analyzing the twitter posts about electronic products like cell phones, computers etc. using Machine Learning approach. By performing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. They presented a new feature vector for classifying the tweets as positive, negative or neutral and extract people's opinion about products [1].

Another research tried to pre-processed the dataset, after that extracted the adjective from the dataset that have substantial meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning algorithms such as Naïve-Bayes, Maximum Entropy and SVM along with the Semantic Orientation based Word-Net which extracts synonyms and relation for the content feature. At the end, they measured the performance of classifier in terms of recall, precision and accuracy [2].

Some researchers had an approach where posted tweets from the Twitter micro-blogging site are subjected to preprocessing and classified based on their emotional content as positive, negative and neutral or irrelevant; and compares the performance of various classifying algorithms based on their precision and recall in such cases. Further, the paper also discusses the applications of this research and its limitations [3].

A number of machine learning like Naïve Bayes and Random Forest models performed sentiment analysis on product review data [8]. Some work in this field included experiments with mood classification on blog posts. One of the researches also deals with review of aspect-based opinion polling from unlabelled free-form textual customer reviews without requiring customers to answer any questions [10].

The tweet retrieval process needs access tokens from the twitter developer site and a piece of code which perform the operation of retrieving those tweets. As the base language used will be java, we choose to implement the Java library called Twitter4J. This library is developed for the twitter API. With Twitter4J library, you can easily integrate your Java application with the Twitter service.

A. Twitter4J has the following features like:

100% Pure Java - works on any Java Platform v5 or later, Android platform and Google App Engine ready, Zero dependency: Additional jars are not required, Built-in OAuth support, already built gzip support.

There are some system requirements that needs to be followed for the Twitter 4J java library to successfully operate. The library supports Windows and Unix Operating systems with Java 1.5 or higher versions installed on it. A java document is also provided in case a user needs to find the method name, syntax, or the root package while implementing the code. To use the java library, the user just needs to add the .jar file to the java application classpath.[4]

III. EXISTING SYSTEM

The existing system works only on the dataset which is constrained to a particular topic. The existing systems also do not determine the measure of impact the results determined can have on the particular field taken into consideration and it does not allow retrieval of data based on the query entered by the user i.e. it has constrained scope. In simple words, it works on static data rather than dynamic data. Unsupervised algorithms like Vector Quantization, are used for data compression, pattern recognition, facial and speech recognition, etc and therefore cannot be used in determining sentiment in twitter data. Apriori algorithm fails to handle large datasets and as a result can generate faulty results.

IV. PROPOSED SYSTEM

In the proposed system, we 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 supervised algorithm on the stored data. The supervised algorithm used in our system is Support Vector Machine (SVM). The results of the algorithms i.e. the sentiment 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.

All rights reserved by *www.ijirst.org*

A. System Architecture

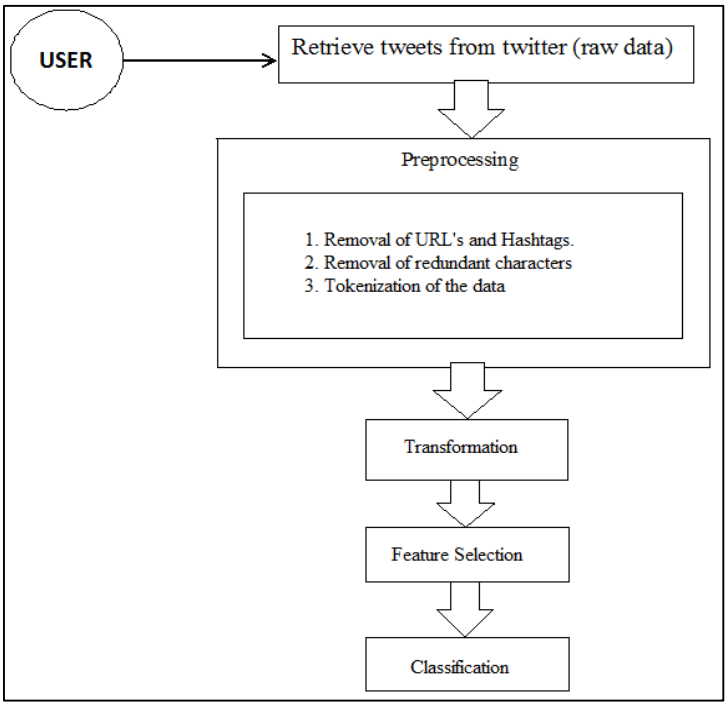

Fig. 1: System architecture

V. COMPARISON

Table – 1
Comparison between existing system and proposed system

Sr. No	*Existing System*	*Proposed System*
1.	*Existing system takes a stored dataset on a particular topic into consideration.*	*Proposed system will gives you the freedom to choose the data of any topic.*
2.	*It fails to determine the impact the results might or will have in the respective field.*	*Here, it gives you the impact the results and statistics will have on the respective field.*
3.	*Existing system does not allow the retrieval of data based on the query entered by user.*	*Proposed system allows retrieval of data based on the query entered by the user.*
4.	*Existing system does not provide accurate feature selection.*	*Proposed system will provide accurate feature selection.*

VI. CONCLUSION

Sentiment analysis has become an important factor in decision making process in a particular field. In this paper we discussed techniques for preprocessing and information retrieval of tweets through twitter. Also we studied about the supervised learning technique: Support Vector Machine for text categorization which can be used to find out the polarity of textual tweet. From study we can conclude that SVM acknowledges some properties of text like High Dimensional feature space, few irrelevant feature, sparse instance vector. The performance of SVM can be evaluated using precision and recall. Different results show that SVM gives good performance on text categorization as compared with ANN. With ability to generalize high dimensional feature space, SVM eliminates need of feature selection.

REFERENCES

[1] Geetika Gautam, Divakar Yadav. (2014). Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. IEEE 2014.
[2] Neethu M S, Rajasree R. Sentiment Analysis in Twitter using Machine Learning Techniques. IEEE 2013.
[3] B. Gokulkrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera,. Opinion Mining and Sentiment Analysis on a Twitter Data Stream. IEEE 2012.
[4] http://twitter4j.org/en/
[5] https://en.wikipedia.org/wiki/Vector_quantization
[6] https://en.wikipedia.org/wiki/Apriori_algorithm
[7] Phillip Tichaona Sumbureru. Analysis of Tweets for Prediction of Indian Stock Markets. IJSR 2013.
[8] Xing Fang, Justin Zhan. Sentiment analysis using product review data. Journal of Big Data 2015.
[9] Gilad Mishne. Experiments with Mood Classification in Blog Posts. Live Journal 2005.
[10] S. A Kanade, S. Shibu and Abhishek Chauhan. Review of Aspect Based Opinion Polling. IJREST 2014.

All rights reserved by *www.ijirst.org*