International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

A review on: Sentiment polarity analysis on Twitter data from different Events

Harshal Kapase[1], Kalyani Galande[2], Tanmay Sonna[3], Deepali Pawar[4], Dipmala Salunke[5]

[1,2,3,4] BE Student & JSPM's Rajarshi Shahu College of Engineering, Pune, India
[5]Professor, Department of IT Engineering, JSPM's Rajarshi Shahu College of Engineering, Maharashtra, India

---

Abstract - *Social media is one of the popular source for information retrieval and also it is being used for sharing day-to-day events of our lives and the incidents which are happening around the world. Twitter can be used as micro-blogging service used to discover events and news in real time from anywhere in the World. Twitter posts are generally short and can be generated constantly, so we can say that they are well-suited sources of streaming data for opinion mining and sentiment polarity detection. Twitter posts on a particular event or subject can help us to know opinions of people about that particular event or subject. Sentiment analysis on twitter posts can help us know how people react to a particular event and how their opinion can change if something unusual happened. Sentiment analysis helps in many business areas to know reviews of an event. While doing any Machine Learning task, the main concern is an accuracy of a model. If the dataset is precise, then we can get a higher accuracy of Machine Learning model. The objective of this paper is to discuss methodology which gives more accuracy of the machine learning task to find out sentiment polarity on Twitter data.*

Key words: Opinion Mining, Sentiment polarity, Dual Sentiment Analysis, Machine Learning, Textual pre-processing

1. INTRODUCTION

For interacting with the world social media is considered a very popular source of information as well as for opinion mining. Opinion mining is a type of natural language processing to track opinions of people about a particular event or subject.

Twitter is the most commonly used social media for expressing opinions on a particular event or an incident [1]. If we consider textual posts on Twitter, they are short and have less size, therefore they are not difficult to process for any experimental purpose. More than 400 million tweets are generated every day and each tweet has maximum 140 characters. A large collection of tweets on a particular event can be used for sentiment analysis to find out people's opinion about that event. If we have a piece of text, then identifying and categorizing opinions expressed in a piece of text by using computational process is Sentiment Analysis. To have a good accuracy of machine learning model one should have the precise collection of a dataset.

If we perform Dual Sentiment Analysis (DSA) on input data of Machine Learning process then we get higher accuracy which is the expected result. DSA helps to improve the accuracy of our model by increasing the data in the dataset. DSA improves accuracy by increasing dataset of original and reverse training reviews. Dual Sentiment Analysis helps in many ways to determine the pole of people towards some particular subject. Sentiment polarity returns the overall opinion of a text or document for one single issue [1]. Sentiment polarity is classifying a given sentence, document or feature into positive, negative or neutral. In the recent years, sentiment polarity helps in many fields like business sectors, government sectors, etc.

For an algorithm to work conveniently and accurately on textual data of Tweets, we need to have exact data of words. As tweets may contain words which are difficult to understand and cannot process further, we need to apply Tweets pre-processing on textual data before sending tweets to sentiment polarity classifier. By using tweets pre-processing one can apply methods like processing texts, classifying, tokenizing, stemming, tagging, parsing, etc on textual data.

2. TECHNIQUES DISCUSSED

The methodological study of opinion mining and sentiment analysis techniques [1] Pravesh Kumar Singh, Mohd Shahid Husain made an attempt to review and evaluate various techniques which are used for opinion mining and sentiment analysis. They have claimed that decision making at individual and organizational level is much more important by studying people's opinion on some topic and what people think about something, so opinion mining can be done by using various opinion-rich resources like reviews, forum discussions, blogs, macro-blogs, Twitter, etc. The views and thoughts of people are subjective figures which express opinions, sentiments, emotional state or evaluation of someone. In this paper, authors have presented different methods for data extraction. They have discussed four methods of Machine Learning-

1. Naïve Bayes Classifier
2. Support Vector Machines
3. Multi-Layer Perceptron

4. Clustering Classifier

In this paper, authors have given benefits and limitations of every method and one can use these methods according to the situation for feature and text extraction. Based on the survey made by authors following table is an example shows the accuracy of different methods using N-gram feature.

	Movie Reviews			Product Reviews		
N-Gram Features	NB	MLP	SVM	NB	MLP	SVM
	75.50	81.05	81.15	62.5	79.2	79.4

According to survey authors claimed, SVM has better accuracy than other methods. After examining four methods in [1], authors came up with the conclusion that Naïve Bayes classifier is best suitable for textual classification, clustering for customer services and SVM for biological reading and interpretation.

A paper Automatic Unsupervised Polarity Detection on a Twitter Data Stream Diego Terrana, Agnese Augello, Giovanni Pilato [2] claimed that a simple and completely automatic system which performs sentiment analysis of tweets of users. Here they used a method which built a corpus by grouping tweets expressing positive and negative polarity through the completely automatic procedure by using only emoticons in the tweets. So they used only emoticons to build a trained model. This method map colloquial expressions with new words, slangs and errors. As they have used emoticons for building a trained model, this method can be applied to any language. The limitation of this method is we need to have tweets which contain emoticons otherwise the method cannot give us an accurate result.

Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li [3] proposed a model of which performs sentiment analysis using Bag of Words (BOW). Here they used Dual Sentiment Analysis (DSA) to solve the problem of polarity shift. Here they have extended the DSA framework into DSA3 which could deal 3-class sentiment classification (positive-negative-neutral) which also takes neutral reviews into consideration. In this paper, they have focussed on creating reversed reviews to assist supervised sentiment classification. This method outperforms the sentiment analysis because as they have used Bag of Words (BOW) which breaks the word order, disrupts the sentiment structures and discards some syntactic information.

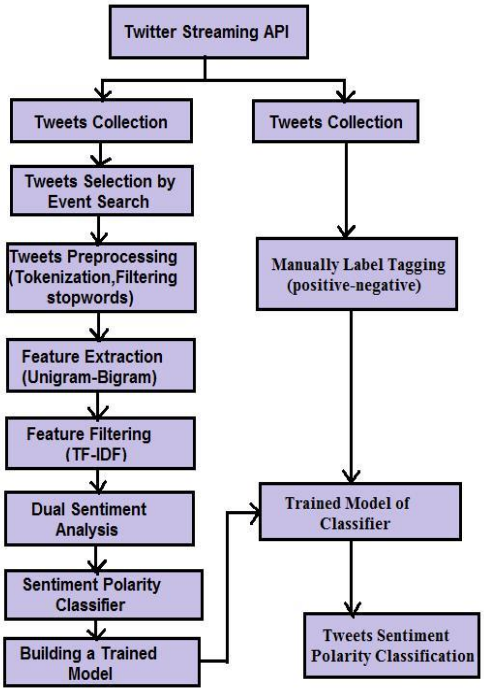
Peiman Barnaghi, John C. Breslin and Parsa Ghaffari [4] proposed a system which provides a positive or negative sentiment on Twitter posts. Twitter streaming API is used for collection of data. The trained model which they have used to perform classification is Bayesian Logistic Regression (BLR). They have used the dataset of FIFA world cup 2014 as their case study. They have trained a sentiment model based on Twitter data using text features. Also, authors have examined major events that occurred during the World cup. The experimental results showed the positive and negative reactions of people towards those major events and how they can change based on the incidents. This kind of sentiment analysis which is done can help players and teams to improve their performance. If we do not have a collection of precise data which is collected from Twitter API, then this model outperforms the accuracy because of a small dataset.

In the paper Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis [5] Geetika Gautam and Divakar Yadav proposed a system which performs classification on customers' reviews, where opinions can be positive, negative or somewhere in between them. They have used Naïve Bayes, Entropy, Support vector machine and semantic analysis (WordNet) algorithms to measure accuracy, precision and recall. The Naïve Bayes method gives better accuracy than the maximum entropy and SVM. The accuracy of the Semantic Analysis is 89.9%. The drawback of this system is that Entropy and SVM do not give better results.

In the paper Application of Machine Learning Techniques to Sentiment Analysis [6] the author has proposed a system which provides a Text analysis framework for twitter data using Apache spark and hence is more flexible, fast and scalable. Decision tree, Support vector machine and Naïve Bayes algorithms are used for sentiment analysis in the proposed framework. The accuracy of the result is 70% to 90%. Decision Tree performance is better than Naive Bayes algorithm. Decision tree performance gives 100% accuracy, precision- recall and F1_score. A limitation is the Naive Bayes does not work as expected on small data set if a dataset is not precise.

A paper Sentiment Analysis on Twitter using Streaming API [7] talks about a system which provides a sentiment analysis on Twitter data with the objective of providing a time-based analytics to the user and real-time sentiment analysis on tweets. The tweets are extracted using Streaming API on a twitter and they are loaded in Hadoop and it is pre-processed using map reduce. Uni-word Naive Bayes Classification is used to classify the tweets. The limitation of this system is, this system use only uni-gram classification and for perfect and accurate results they needed to use N-gram classification.

3. SENTIMENT ANALYSIS ON TWITTER DATA SYSTEM OVERVIEW



1. Tweets collection:

This step comprises of collecting the required number of tweets from Twitter API for the creation of training set. The data is in JSON format as a set of documents.

2. Tweets pre-processing:

To find out sentiment analysis the major step is filtering out all the noise and meaningless symbols that do not contribute to a tweet sentiment from the original text. The processes like tokenization, removing stop-words, stemming, etc also perform during tweets pre-processing.

3. Feature extraction and feature filtering:

This step comprises of selecting useful list of words as a feature of text and remove a large number of words that do not contribute to the text sentiment. The concept of N-grams is used where a set of sequential words is used for finding out the frequency of words. After these processes, we apply algorithm of Dual Sentiment Analysis to increase the accuracy of a model.

4. Trained model of classifier:

After processing all the data, the final result is a trained model which is applied to test data for checking the accuracy of a model. A trained model is built by using a classification method.

4. CONCLUSION AND FUTURE WORK

The project "Sentiment analysis on Twitter data" aims to detect whether a tweet is positive or negative, it can help us to know reviews of people about any subject/topic. It can also help us to know whether the situation in some disaster area is critical or not. So by using those results, it would be easy for social charities and government to take many decisions for the sake of people. Also, customer reviews help organizations to improve in some particular areas and also they would have come to know the appreciation of customers for some product/idea. Also, our main concern in Machine Learning Algorithm is the accuracy of the result, so our proposed system introduces Dual Sentiment Analysis (DSA) which gives us large dataset. As a result, the accuracy of our model also increases which is our main concern.

As we have seen that this model can be used to detect whether the tweet is positive or negative, if we fetch the tweets regarding disaster place and if we find out that most of the tweets are showing "negative" that means the situation in that area is critical, so as the future scope we can design an algorithm which can fetch all the "Negative" labelled tweets and will find out the major issues in that area and will directly inform charities about those major issues.

REFERENCES

[1] Pravesh Kumar Singh, Mohd Shahid Husain, "Methodological study of opinion mining and sentiment analysis techniques", International Journal on Soft Computing (IJSC) Vol. 5, No. 1, February 2014, DOI: 10.5121/ijsc.2014.5102

[2] Diego Terrana, Agnese Augello, Giovanni Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream", 2014 IEEE International Conference on Semantic Computing, 978-1-4799-4003-5/14 $31.00 © 2014 IEEE

[3] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", 1041-4347 (c) 2015 IEEE.

[4] Peiman Barnaghi, John C. Breslin and Parsa Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", 2016 IEEE Second International Conference on Big Data Computing Service and Applications, 978-1-5090-2251-9/16 $31.00 © 2016 IEEE

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations and trends in information retrieval, vol. 2, pp. 1-135, 2008.

[6] A. Bifet and E. Frank, "Sentiment knowledge discovery in Twitter streaming data", in Discovery Science, 2010, pp. 1-15.

[7] Diego Terrana, Agnese Augello, Giovanni Pilato, "Facebook Users Relationships Analysis based on Sentiment Classification", 2014 IEEE International Conference on Semantic Computing, 978-1-4799-4003-5/14 $31.00 © 2014 IEEE.

[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations and trends in information retrieval, vol. 2, pp. 1-135, 2008.

[9] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data", in Discovery Science, 2010, pp. 1-15.

[10] Diego Terrana, Agnese Augello, Giovanni Pilato, "Facebook Users Relationships Analysis based on Sentiment Classification", 2014 IEEE International Conference on Semantic Computing, 978-1-4799-4003-5/14 $31.00 © 2014 IEEE

[11] Addlight Mukwazvure, K.P Supreethi, "A Hybrid Approach to Sentiment Analysis of News Comments", 978-1-4673-7231-2/15/$31.00 ©2015 IEEE

[12] C. Huang S. Li, Y. Lee, Y. Chen and G. Zhou (2015), "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis ", in Proc. Annu. Meeting Assoc. Compute Linguistics.

[13] G.Grace Ranjitham, S.Mohana, B.Vinothini, "Sentiment Analysis For Two Sides Of Reviews Using Dual Prediction Algorithm", All Rights Reserved © 2016 IJARTET

[14] M. Trupthi, Suresh Pabboju, G. Narasimha (2017),"Sentiment Analysis on Twitter using streaming API", 978-1-5090-1560-3/17 $31.00 © 2017 IEEE DOI 10.1109/IACC.2017.177