Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Opinion Detection System for Twitter using Machine Learning Approach

Priyanka Gupta*, Rajiv Kumar
Department of Computer Science and Engineering, Sharda University,
Greater Noida, India

Abstract— Opinion analysis is a process of computationally recognizing and classifying beliefs expressed in a piece of document, mainly in order to figure out whether the author's perspective towards particular theme or topic, product etc. is negative, positive or neutral. Today micro-blogging has become a very crucial and important transmission link among web users. Millions of messages are transferred through social media platforms like twitter, Facebook etc. and web user's of those messages exchange their opinions about their life and also on various current topics. Social media platform like Twitter provides easy accessibility and word limit of 140 characters to express their opinions bend the web users to shift from conventional communication tools to web blogging services. Therefore, web blogging platforms have become a valuable source of information which can be efficiently used in various fields such as marketing, social studies, spreading social awareness, research etc. In this paper, an opinion analysis approach has been applied to a group of tweets related to a natural disaster happened in Nepal; a prototype is designed and trained using bagging classification techniques. Thus, result provides beneficial information which will help the experts to plan their operations and hence alleviate the operations of managing such disaster situations.

Keywords— Opinion Analysis, Machine Learning Approach, Web data Analysis, User Generated Content.

I. INTRODUCTION

Computer science has various sub-disciplines from which Machine learning is one of the sub-discipline [1] that developed from the learning of pattern recognition and methodical learning in artificial intelligence. Machine learning traverses the study and fabrication of algorithms that can retain from and make forecasting on data.

Opinion analysis is a machine learning approach in which system analyzes and categorizes the human's opinions, emotions, sentiments etc about some topic or theme which are conveyed in various format like text or speech. The textual document available in the internet is increasing day by day. For enhancing the sales of a commodity and to improve the buyers' satisfaction, generally on-line shopping websites provide the chances to buyers' to write opinion about commodity. These opinions are huge in number and to draw the overall opinion polarity from all of them, opinion analysis can be utilized. Manually analyzing such huge number of reviews is practically impossible and time consuming. Therefore automatic approach of a machine has notable role in resolving this hard problem. The major issue of the field of Opinion analysis and Sentiment mining lies in recognizing the emotions conveyed in these texts.

Opinion analysis is highly Domain centered approach. This means that a solution you have evolved for one domain (e.g. Movies) will not mechanically work on other domains (e.g. cell-phones). Figure 1 shows an example of tweets with their corresponding sentiment values.

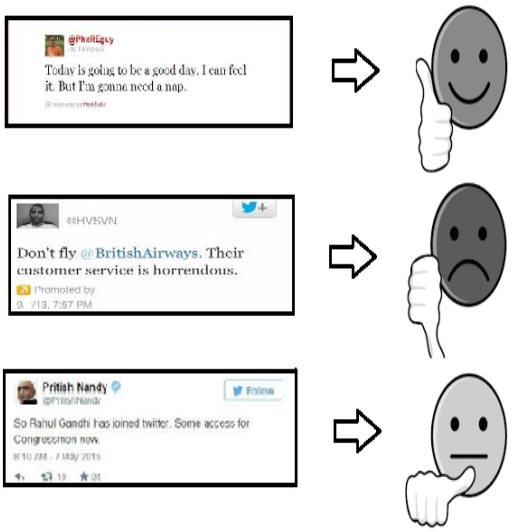

Figure 1: Tweets with their corresponding Sentiment Value

Motivation for Opinion Analysis is two-step approach. Both buyers and manufacturers highly value "customer's views" about products and services. Thus, Opinion Analysis has seen a remarkable effort from industry as well as academics.

1. The Buyer's Perspective

While taking a resolution it is very crucial for us to know the belief of the people around us. Earlier this set used to be short, with a few trusted friends and family supporters. But, now with the emergence of Internet we see people indicating their beliefs in blogs and forums. These are now frequently read by people who seek a review about a particular unit (product, movie etc.). Thus, there is a plethora of beliefs available on the Internet.

From a buyers point of view withdrawing opinions about a specific entity is very important. Trying to go through such a huge amount of information to grasp the general opinion is unfeasible for users just by the sheer capacity of this data. Hence, the necessity of a system that distinguishes between good reviews and bad reviews. Further, labeling these logs with their opinion would provide a concise summary to the readers about the common opinion regarding an entity.

2. The Manufacturer's Perspective

With the ignition of Web 2.0 platforms such as blogs, discussion forums, etc., buyers have at their disposal, a platform to express their brand experiences and views, positive or negative regarding any commodity or service. According to Pang and Lee, 2008 these buyer voices can wield extensive influence in shaping the sentiment of other consumers and, ultimately, their brand fidelities, their investment decisions, and their own brand endorsement [7].

Since the buyers have commenced using the power of the web to dilate their horizons, there has been a stream of review sites and blogs, where users can recognize a product or service's advantages and disadvantages. These reviews thus shape the future of the commodity or the service. The manufacturers require a system that can recognize trends in public reviews and use them to upgrade their product or service and also recognize the requirements of the future.

3. The Communities' Perspective

Recently, certain affairs, which affected administration, have been triggered using the web. The social grids are being used to bring jointly people so as to assemble mass gatherings and oppose oppression.

On the darker side, the social grids are being used to suggest people against an ethnic group or set of people, which has concluded in a serious loss of life. Thus, there is a requirement for Sentiment Analysis systems that can recognize such phenomena and diminish them if needed.

This research paper is structured mainly in four Partitions: Partition II gives quick overview about the recent development happened in Opinion Analysis Domain and a comparative analysis in Different domain with their performance analysis is mentioned, Partition III discusses the proposed architecture for opinion analysis, and lastly, Partition IV deduces the paper.

II. RELATED WORK

Past work in that area consist of Turney (Turney et al, 2002) and Pang (Pang et al, 2002) who applied different procedure for detecting the polarity of commodity reviews and movie reviews respectively. This piece of work is at the document level. One can also classify a document's polarity on a multi-way scale, which was attempted by Pang and Snyder (Snyder et al, 2007) among others: Bo and Lilian (Lilian et al, 2005) expanded the basic task of classifying a movie review as either positive or negative to predicting star ratings on either a 3 or a 4 star scale, while Snyder performed an in-depth analysis of restaurant reviews, predicting ratings for various aspects of the given restaurant, such as the food and atmosphere (on a five-star scale). Moreover it is also proven that particular classifiers such as the Max Entropy (Vryniotis et al, 2013) and the SVMs (Koppel et al, 2006) can profit from the introduction of neutral class and improve the overall accuracy of the classification.

B.Pang (B. Pang et. all 2004), the instigators present a method of subjectivity recognition for sentiment analysis. This is essential because the unrelated data from the reviews could be eradicated. This eradicates the processing overheads of a huge amount of textual data. The procedure they recommend is using minimum cuts to yield subjective extracts from the text. The work has been concentrated in the sentence level subjectivity dilation.

A classification technique using Naive Bayesian classifiers used in (J. Wiebe et. all, 2005). They show the results of evolving subjectivity classifiers using un-annotated texts for training. In this piece of work of learning Subjective and Objective sentences, the procedure automatically generate training data. This is executed by a Rule-based method. The rule-based subjective classifier categorizes a sentence as subjective if it includes two or more strong subjective clues. They use Subjective Precision, Subjective Recall, Subjective F measure, Objective Precision, Objective Recall and Objective F measure for the analysis. They also implement a self-training procedure for the system.

A remarkable approach was used by B. B.Khairullah Khan (B. B.Khairullah Khan et. all, 2010) who uses a sentence level opinion analysis. The word level feature dilation is done using Naive Bayesian Classifier. The semantic alignment of the individual sentences is recovered from the contextual information. This machine learning method on average claims a precision rate of 83%. For classifying and inspecting of the sentiment from the reviews, machine learning and lexical contextual statistics are used. The paper focuses on sentence level to examine whether the sentences are objective or subjective and to categorize the polarity of the sentences to positive or negative sentiment.

R. B. W. N. Jeonghee Yi (R. B. W. N. Jeonghee Yi, T Nasukawa, 2003) the instigators present a method for opinion mining which depends on natural language processing approaches. The work is fulfilled by the sentiment lexicon and a pattern database. The two feature selection algorithms conferred in this work are grounded on mixture model and the likelihood ratio. They suggest a sentiment pattern based analysis for the opinion classification work.

T. Nasukawa (T. Nasukawa et. all, 2003), who uses an opinion analysis approach to dilate sentiments associated with polarities of positive or negative for particular subjects from a document is executed. This is in variation of classifying the whole document into positive or negative. In order to recognize sentiment expressions and to examine their semantic association with the subject term, natural language processing plays a crucial role. The procedure identifies the subjects in the beliefs sentences and link opinions to these subjects.

Another notable work is the implementation of both Natural Language understanding and Creation in Sentiment analysis (M. Hu et. all, 2004). A couple of algorithms to find and predict the orientation of beliefs are mentioned in this research work. In their routine there is an inspection database that stores the opinionated texts. The procedure then finds chronic features that many people have expressed their beliefs on. After that, the opinion words are dilated using the resulting chronic features, and semantic alignment of the opinion words are recognized with the help of WordNet. The routine then searches those infrequent features. The alignment of each opinion sentence is recognized and a final text summary is created in this work. The part of speech (POS) tagging from natural language processing is used to find opinion characteristics. The result of the above paper is a text summary of sentiment. Thus Summarization of text is also executed as a subsystem. But this summarization work is truly dependent on the characteristics and hence is far from the mechanical summarization work in the field of NLP. The paper suggests a method by consuming the adjective synonym set and antonym set in WordNet to forecast the semantic orientations of adjectives. The paper also mentioned the need of pronoun resolution in sentiment mining even though it is not addressed.

Support Vector Machine (SVM) Classifier is made employ of in (T. Mullen and N. Collier, 2004,p. 412-418). The technique highlights the use of a collection of diverse information sources, and SVMs furnish the ideal tool to conduct these sources together. The procedures are used to allocate values to selected words and phrases, and guide them together to develop a model for the classification of texts. In this paper, the sentiment alignment of a phrase is determined grounded upon the PMI (phrase's point wise mutual information) with the words like good and poor. Semantic values of phrases and words inside a text are utilized to add to features for SVM training. Association of SVMs using these features in co-occurrence with SVMs grounded on unigrams and lemmatized uni-grams are a different method according to them.

Various methods are available in order to solve the problem of opinion analysis but it is very complex to declare which method will give optimal results as every method has its own advantages and disadvantages.

III. EXPERIMENT

Figure 3 shows the procedure used in order to identify opinions from twitter in calamity situation. As mention in figure 3: Opinion recognition is conducted in 8 levels which are as follows: Data collection, Data preprocessing, Data Analysis, Model Creation, Feature extraction & reduction, Training & model evaluation further concludes into Classification Results and stored in data storage.

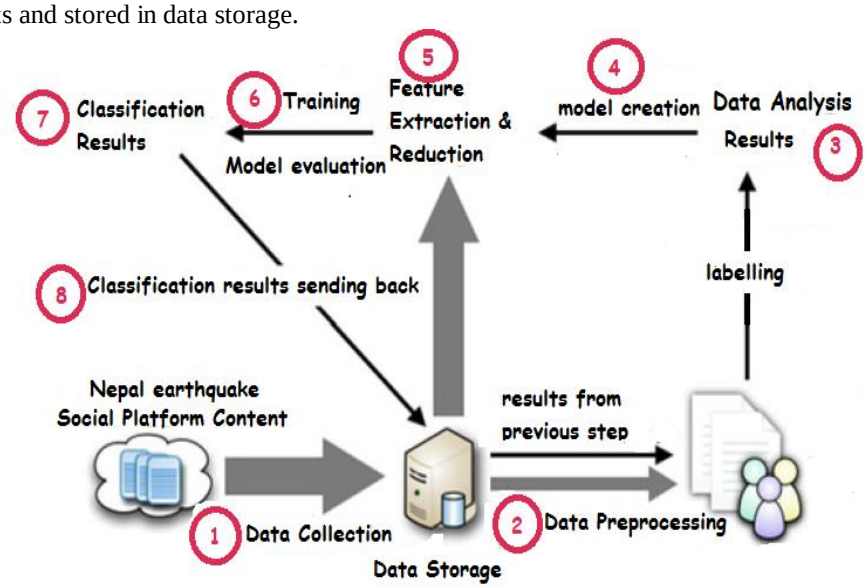

Figure 3: Procedure for Opinion Analysis for Tweets

A. Data Collection:

In this level data is gathered from social platform i.e. Twitter for the duration of 2 months by using twitter Application program interface [9] using hashtags #earthquake, #nepal and #nepal earthquake. Nepal earthquake came recently and destructed the entire city. 10,000 tweets were gathered from which 3,000 were found to be impertinent. Therefore, only 7,000 tweets were taken into consideration for the proposed methodology [12].

B. Data Pre-Processing:

In this level the gathered data is pre-processed as it includes vast amount of irrelevant data which is in the form of misspelling, repetition of alphabets, slang words etc. Therefore it is foremost to pre-process the gathered data in order to make it consistent & eliminate the irrelevant words which do not impart any opinion in the drafted text. The Nepal earthquake dataset also includes huge tweets which are in Hindi i.e. translated into English by using Google Translate API. So, following steps are taken in order to pre-process the dataset [10]:

1. In order to make the data set consistent all the tweets are transformed into lower case.
2. All twitter user names and internet addresses mentioned in the twitter tweets are substituted with a constant string.
3. All Hash tags are eliminated that is all #words are substituted with word, e.g. # Nepal earthquake are substituted with Nepalearthquakes.
4. All irrelevant white spaces, punctuations, special symbols and alpha-numeric characters are eliminated.
5. Common Stop words are eliminated.
6. All the repeating characters mentioned in a word are eliminated. E.g. verrry is substituted with very.

C. Data Analysis:

In this level the pre-processed data is further labeled with three major emotion classes: positive, neutral & negative based on Paul Ekman's Basic Sentiments Theory [11] and Labeling Principles. After studying the emotions expressed in 7,000 related tweets 2,500 tweets are labeled as positive, 3000 tweets are labeled as negative and rest are labeled as neutral as shown in table 1.

Table 1: Data Analysis Outcomes

Gathered Tweets	Negative	Positive	Neutral
7,000	3000	2500	1500

D. Model Creation:

A prototype is designed which is a fusion of N gram and stemming technique.

E. Feature Extraction & Reduction:

For attribute reduction String to word vector filter is utilized that is to elect a subset of important attributes to be utilized for prototype construction.

F. Model Evaluation & Training:

90% of the above labeled twitter tweets i.e. 6,300 are utilized as a training data set and 10% i.e. 700 tweets are utilized as a testing dataset. How dataset is bifurcated into training and testing datasets is also mentioned in Table 2.

Table 2: Data bifurcation into Training &Testing Datasets

Dataset	Number of tweets
Training Dataset	6,300
Testing Dataset	700

G. Classifiers:

Bagging: Bagging is one of the ensemble techniques which are used to enhance the accuracy of the system. Bagging is basically a bootstrap aggregation algorithm which is used for decreasing the variance of our predictions. It generates additional information for training from our original data sets by using group of repetitions to create multiple sets of same size. It also tunes our predictions which will help in producing expected results.

The Four Crucial Measures that has been used to analyse the performance of classification models are mentioned below:

$$\text{Accuracy (a)} = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \quad \text{——— (1)}$$

$$\text{Precision (p)} = \frac{Tp}{Tp + Fp} \quad \text{——— (2)}$$

$$\text{Recall (r)} = \frac{Tp}{Tp + Fn} \quad \text{——— (3)}$$

$$\text{F1 Score} = \frac{2*p*r}{p + r} \quad \text{——— (4)}$$

Table 3: Generalized Contingency Matrix

	Not Actual Condition	Actual Condition
Predicted Condition	False Positive (Fp)	True positive(Tp)
Not Predicted Condition	True negative(Tn)	False negative(Fn)

H. Classifications Outcomes:

All-Inclusive accuracy of 74.4053% is obtained on the Nepal Earthquake Dataset using Multinomial Naïve Bayes Classifier. Table 4 shows the values of other evaluation parameters which are calculated for each class separately.

Table 4: Evaluation parameters with Multinomial naïve Bayes

	Negative	Neutral	Positive
Precision	0.732	0.286	0.763
Recall	0.756	0.097	0.768
F1-Score	0.744	0.145	0.766

An-inclusive accuracy of 67.9295% is obtained on the Nepal Earthquake Dataset using random tree. Table 6 exhibits the values of other evaluation parameters which are calculated for each class separately.

Table 5: Evaluation parameters with Random Tree

	Negative	Neutral	Positive
Precision	0.661	0.118	0.706
Recall	0.712	0.032	0.684
F1-Score	0.686	0.051	0.695

An-inclusive accuracy of 71.2775% is obtained on the Nepal Earthquake Dataset using Support vector machine. Table 6 exhibits the values of other evaluation parameters which are calculated for each class separately.

Table 6: Evaluation parameters with SVM

	Negative	Neutral	Positive
Negative	0.707	0.143	0.725
Neutral	0.719	0.032	0.744
Positive	0.713	0.053	0.734

IV. CONCLUSION

A novel approach has been presented in this research paper in disaster domain for viewer's opinion detection. It provides advantageous information that can be utilized by both private & government firms to administer such circumstances in a better and in effective manner. A baseline prototype is developed using Bagging ensemble technique which are trained on the characteristics like Stemming & N-gram. Further we tested the data set. Bagging technique outperforms best using Multinomial naïve bayes classifier with overall accuracy of 74.4053%. For future we aim to implement other machine learning (supervised) approaches like Neural Network, clustering to extract a comparison study between them and to identity the most suitable characteristics fusion and classifiers for Opinion recognition in this fresh domain of Natural calamity.

REFERENCES

[1] Kohavi, R. & Provost, F. 1998. Glossary Of Terms. Machine Learning, 30, 271-274.

[2] Peter, D. Year. Turney: Thumbs Up Or Thumbs Down? Semantic Orientation Applied To Unsupervised Classification Of Reviews. *In:* Proceedings Of 40th Annual Meeting Of The Association For Computational Linguistics, 2002. 417-424.

[3] Li, G. & Liu, F. Year. A Clustering-Based Approach On Sentiment Analysis. Intelligent Systems And Knowledge Engineering (Iske). *In:* 2010 International Conference On. Ieee, 2010.

[4] Hu, M. & Liu, B. Year. Mining And Summarizing Customer Reviews. *In:* Proceedings Of The Tenth Acm Sigkdd International Conference On Knowledge Discovery And Data Mining, 2004. Acm, 168-177.

[5] Baharudin, B. Year. Sentence Based Sentiment Classification From Online Customer Reviews. *In:* Proceedings Of The 8th International Conference On Frontiers Of Information Technology, 2010. Acm, 25.

[6] Nasukawa, T. & Yi, J. Year. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In: Proceedings Of The 2nd International Conference On Knowledge Capture, 2003. Acm, 70-77.

[7] P. H. Theresa Wilson, Janyce Wiebe, "Proceedings Of Human Language Technology Conference And Conference On Empirical Methods In Natural Language Processing," Association For Computational Linguistics, 2005, 347354.

[8] Hasan, S. S. & Adjeroh, D. Year. Proximity-Based Sentiment Analysis. In: Applications Of Digital Information And Web Technologies (Icadiwt), 2011 Fourth International Conference On The, 2011. Ieee, 106-111.

[9] Dev.Twitter.Com,. "Twitter Developers". N.P., 2014. Web, 24 Dec. 2014

[10] Pak, A. & Paroubek, P. Year. Twitter As A Corpus For Sentiment Analysis And Opinion Mining. *In:* Lrec, 2010. 1320-1326.

[11] Ekman, P. 1992. Are There Basic Emotions? *Psychological Review,* 99**,** 550-553.

[12] Kaur, H. J. & Kumar, R. Year. Sentiment Analysis From Social Media In Crisis Situations. *In:* Computing, Communication & Automation (Iccca), 2015 International Conference On, 2015. Ieee, 251-256.

