A Review on Sentiment Analysis using Twitter during Crisis

Anuradha Pawar[1] Sugandha Singh[2]
[1]Research Scholar [2]Professor & Head & Department
[1,2]Department of Computer Science & Engineering
[1,2]PDM College of Engineering, India

Abstract— This paper focuses on the popular micro-blogging platform such as Twitter through which individuals share information about any topic across the world. They also share crisis-related information during the mass emergency through Twitter. These tweets are used for sentiment analysis. We will discuss an approach in which tweets are fetched from Twitter's streaming API. The tweets are manually tagged according to respective sentiments i.e positive, negative or neutral. Use of available preprocessing technique in Weka proved better results. After preprocessing Naïve Bayes classifier is applied.

Key words: Twitter, Sentiment Analysis, Weka, Naïve Bayes Classifier

I. INTRODUCTION

Micro-blogging sites such as Twitter, Facebook, and LinkedIn are sites where people share their opinions [3]. Our focus for this review paper is on Twitter, which allows its users to share short messages called Tweets which are of 160 characters or less [1]. Messages that people share are saved in their profiles and also forwarded to others in their circle of friends [4]. The information can be kept private or shared among the public. Therefore, Twitter provides a rich source of information during the crisis for sentiment analysis [6]. When people get to know about the disaster occurrence or forecasting, they start uploading their opinions on social media. Sentiment analysis is a text classification problem which extracts information present in the text. Some opinion represents sentiments and some do not represent any sentiment. Sentiment can be categorized into positive, negative or neutral [7]. Therefore, sentiment provides a good source of information to find out the severity of disaster and better situational awareness to the people.

On the other hand, a large number of organizations have a high demand for mechanisms that will automatically harness the volume of data generated by the people and assist them to evaluate public opinion regarding the topic of interest [2]. Examples of organizations where sentiment analysis in disaster management is used are the American Red Cross Digital Operations Center powered by Dell, the European Union security research project Alert4All [5]. Towards this direction, Twitter provides most promises results to provide relevant information.

In view of above, the data preprocessing is important step in sentiment analysis which increases the correctly classified instances by selecting appropriate preprocessing technique [3].In this paper, we will describe the methodology for collecting crisis related tweets and tagging manually with respective sentiments. We will also describe the effect of preprocessing techniques.

II. LITERATURE REVIEW

Sentiment analysis has been topic of interest for a researcher from last decade. One reason is the growing amount of opinion-rich text messages made available through the social networking platform [5]. Another reason for increased interest is advanced in the field of machine learning and natural language processing.

A survey of various techniques suggested for opinion mining and sentiment analysis is presented in [10]. A seminal work on the use of machine learning for sentiment analysis is showing that good performance can be achieved for the problem of classifying movie reviews as either positive or negative using balanced dataset.

In [8], the author extended the comparison of sentiment polarity classification techniques for tweets. They also used the combination in the compared set and used manually annotated tweets for evaluation of classifiers. As the polarity of tweets used is negative and positive, neutral sentiments cannot be classified.

In [9] using various preprocessing techniques and applying various selection techniques to the Naïve Bayes classifier, they were able to achieve good performance by using training set.

Microblogs such as Twitter face challenges in sentiment analysis since messages are short i.e not more than 140 characters and may contain sarcasm and may contain sarcasm and slang [6].

Social media monitoring techniques for collecting a large number of tweets during the crisis and classifying them with machine learning algorithm has become a popular topic within the crisis response and management domains [11].

In [10], the authors presented an analysis of various parameter settings for selected classifier: Decision Trees, Support Vector machine and Naïve Bayes. They used ngrams of normalized words as features and observed the results for positive, negative or neutral. They made their experiments in the Spanish language for topic cell phones and Mexican presidential elections using balanced and unbalanced datasets. In [1] authors proposed an annotated process of analyzing users emotions and geographic distribution of disasters using tweets from twitter. However, after a thorough investigation in the related scientific literature, we came up with the result that the effect of preprocessing techniques should be analyzed.

III. METHODOLOGY

The main aim of this review paper is to evaluate the role of preprocessing techniques for sentiment analysis. Hence, we examine the performance of machine learning classifiers using various preprocessing techniques on Twitter dataset retrieved through twitter streaming API.

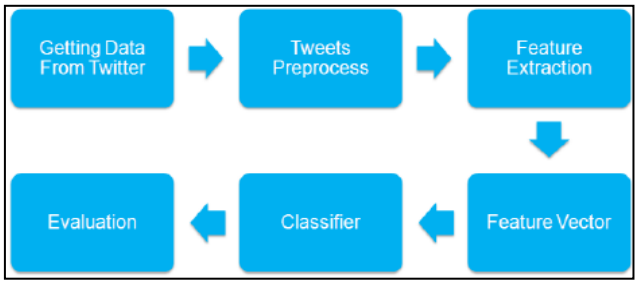

Fig. 1: Methodology for Evaluating Result

A. *Collecting Tweets*

The first step in our methodology was to collect a large set of crisis-related tweets. For this, we used Python package tweet stream to extract tweets related to the Alberta Flood 2013.The tweet stream fetches the tweets from Twitter's streaming API in real-time. The streaming API only gives access to random sample of the total volume of tweets sent at any given moment.

B. *Annotation Process*

After an initial manual review of collected tweets, we discovered that large portion of the tweets belonged to category neutral. So we gave a category neutral also other than positive and negative. By default, Weka accepts dataset in Attribute Relation File Format for data analysis. So, we have to convert tweets dataset into ARFF format.

C. *Preprocessing Technique*

Preprocessing is the most important step for sentiment analysis. To perform preprocessing in Weka, we use the String to Word Vector filter. This filter allows using the following configuration:

1) TF-IDF Weighting Scheme

It is a standard approach for feature vector construction. TF-IDF stands for "term frequency-inverse document frequency" and reflects how important a word in the dataset.

2) Stemming

Stemming algorithms work by removing the suffix of the word, according to some grammatical rules.

3) Stop-Words Removal

It is a technique that eliminates the frequent used words which are meaningless and useless for the text classification. This reduces the corpus size without losing important information.

4) Tokenization

This setting splits the document into words, constructing a word vector known as a bag of words. We propose N Gram Tokenizer to compare word unigram, bigram and 1-to-3-gram.

The above preprocessing generates a huge number of attributes among which many are not relevant:

5) Feature Selection

It is a process by which the number of attributes is decreased into a better subset which can bring the highest accuracy. The benefits of performing this option on the data are the limitation of overfitting, the improvement of accuracy and the reduction in training time. Feature Selection methods can be classified as Filter and Wrappers. Filters are based on statistical tests such Infogain, Chi-square and CFS while Wrappers use a learning algorithm to report the optimal subset of feature. For this task, Weka provides the Attribute Selection filter which allows to choosing an attribute evaluation method and a search strategy. In this paper we examine three options:

6) No Filter Applied

We use all attribute created by String to Word Vector filter.

7) Info Gain Attribute Eval

It evaluates the worth of an attribute by measuring the information gain with respect to the class and we set the Ranker search method.

8) Classification Attribute Eval

It evaluates the worth of an attribute by using a user-specified classifier. We choose Random Forest as the classifier and set the Ranker search method to select the top 70% attributes.

D. *Classifier*

Well, known classifier, namely Naïve Bayes has been used in order to evaluate the performance depending upon the preprocessing methods applied on the dataset. This classifier comprises of most representative machine learning algorithm which is provided in Weka. Naïve Bayes algorithm is based on Bayesian Rule which is as follows:

$$P(C|X) = P(X|C)P(C)/P(X)$$

Where we could say that:

Posterior = Likelihood x Prior/ Evidence

- P(C|X) is the posterior probability which represents the degree to which believe a given model accurately describes the situation.
- P(C) is the prior probability of class which describes the degree to which we believe the model accurately describes the reality based on all of our prior information.
- P(X|C) is the likelihood describes how well the model predicts the data.
- P(X) is the prior probability of predictor.

IV. EXPERIMENTAL RESULT

We used balanced dataset of Alberta Flood 2013 for sentiment analysis in Weka. The positive, negative and neutral sentiments are shown in the figure 2.

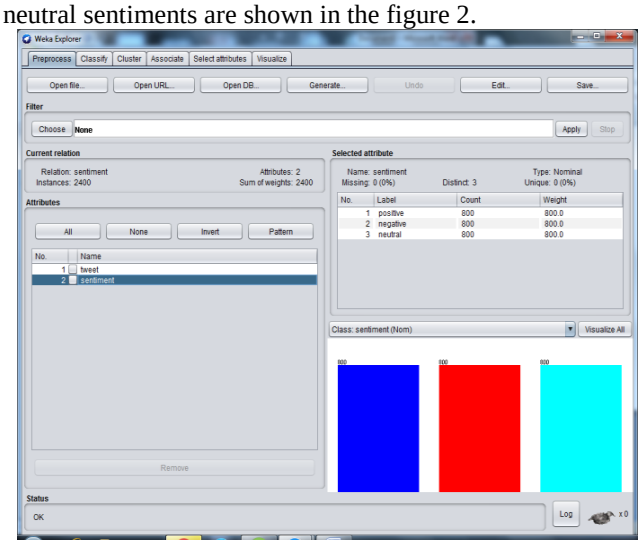

Fig. 2: Balanced Dataset

In order to specify the optimal settings of the preprocessing techniques and classifiers, we conducted variety of experiments that returned more accurate results.

The first experiment contained following techniques listed in table I

Preprocessing Technique	Applied options
Stemming	Snowball Stemmer
Stop-words removal	Rainbow
Tokenization	WordTokenizer

Table 1: Preprocessing Techniques used in Experiment 1

Using the above preprocessing techniques we obtained a good performance using Multinomial Naïve Bayes text classifier. We tried more experiments to increase the performance of the classifier by using following techniques listed in table II.

Preprocessing Technique	Applied options
Stemming	Lovin Stemmer
Stop-words removal	MultiStopwords
Tokenization	NGramtokenizer

Table 2: Preprocessing Techniques used in Experiment 2

Using the above techniques listed in table 2, Naïve bayes Multinomial text classifier gave better performance than techniques listed in table 1.

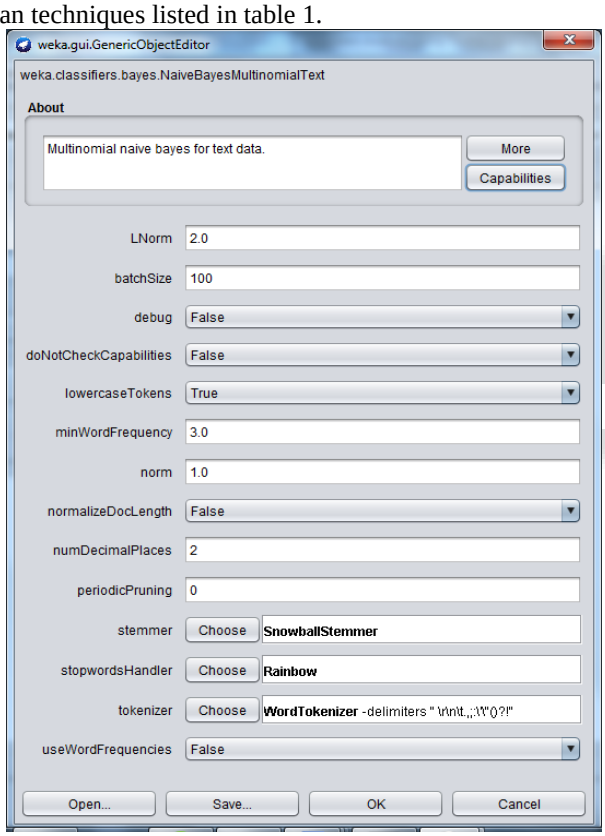

Fig. 3: Preprocessing Techniques Applied in Experiment 1

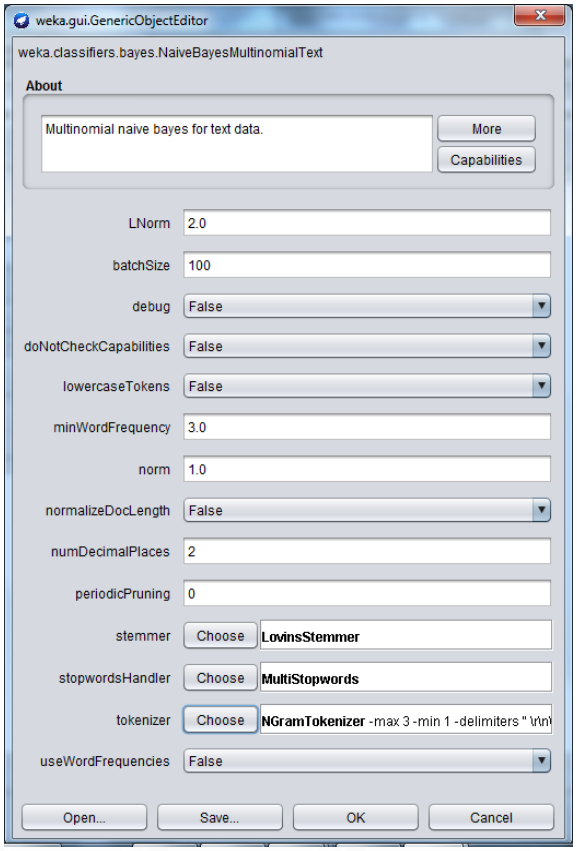

Fig. 4: Preprocessing Techniques Applied in Experiment 2

V. Conclusion & Future Work

For the classification of sentiments, a primary phase of text preprocessing operations and feature extraction is essential. The preprocessing operations affect the quality of the classification, for this we performed various experiments to obtain better performance. Using our dataset use of Lovin stemmer, NGramtokenizer, MultinomialStopwords increased the performance of classifier.

However, it is needed to explore more preprocessing techniques to find optimal solution. Additionally, the sarcasm is needed to be detected in the tweets for better classification of tweets.

References

[1] Himanshu Shekhar, Shankar Setty "Disaster Analysis Through Tweets", International Conference on Advances in Computing, Communications and Informatics (ICACCI),IEEE Aug 2015

[2] Akrivi Krouska, Christos Troussas, Maria Virvou "The effect of preprocessing techniques on Twitter Sentiment Analysis",7th International Conference on Information, Intelligence, Systems & Applications (IISA),IEEE July 2016

[3] Umadevi V, "Sentiment Analysis using Weka", International Journal of Engineering Trends and Technology-Volume 18 number 4-Dec 2014

[4] Shruti Wakade, Chandra Shekar, Kathy J.Liszka and Chien-Chung Chan, "Text Mining for Sentiment Analysis of Twitter Data"

[5] Joel Brynielson, Fredrik Jonsson, Carl Jonsson and Anders Westling, "Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises", Security informatics, a SpringerOpen Journal
[6] Pooja kumari, Shikha Singh, Devika More, Dakshata Talpade, "Sentiment Analysis of Tweets",International Journal of Science Technology & Engineering, Volume 1,Issue 10, April 2015
[7] Ankit Pradeep Patel, Ankit Vithalbhai Patel, Sanjaykumar Ghanshyambhai Butani, Prashant B.Sawant, "Literature Survey on Sentiment Analysis of Twitter Data using Machine Learning Approach", International Journal of Innovative Research in Science & Technology, Volume-3, Issue-10, March 2017
[8] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou,"Comparing methods for twitter sentiment analysis," KDIR 2014-Proceedings of the Int. Conf. on Knowledge Discovery and Information Retrieval, pp. 225-232, 2014
[9] S. Fouzia Sayeedunnissa, A. R. Hussain, and M. A. Hameed,"Supervised opinion mining of social network data using a bag-of-words approach on the cloud", 7th International Conference on Bio-Inspired Computing: Theories and Application, 2013.
[10] G.Sidorov, S.Miranda-Jimenez,F.Viveros-Jimenez and J.Gordon, "Empirical study of machine learning based approach for opinion mining in tweets", 11th Mexican International Conference on Artificial Intelligence,MICAI,2012
[11] B Pang, L Lee, Opinion mining and sentiment analysis. Fondations Trends Inf Retrieval. 2(1-2), 1-135(2008). Doi:10.1561/1500000011
[12] J Yin, A Lampert, MA Camerin, B Robinson, R Power, Using social media to enhance emergency situation awareness IEEE Intell Syst. 27(6), 52-59(2012). Doi:10.1109/MIS.2012.6

