Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 5, May 2018, pg.76 – 82

Sentiment Analysis in Disaster Management using Tweets

Anuradha Pawar[1], Sugandha Singh[2]

[1]Research Scholar (CSE Deptt.), PDM College of Engineering, Bahadurgarh, Haryana, India
[2]Professor & Head (CSE Deptt.), PDM College of Engineering, Bahadurgarh, Haryana, India
[1] annupawar1991@gmail.com; [2] sugandhasinghhooda@gmail.com

Abstract—Social media platform such as Twitter is becoming a source for event-based early warning systems. These event-based tweets act as input to crisis management organizations for decision making. In this paper dataset containing tweets of the sandy hurricane is used for sentiment analysis. Sentiment analysis is used to analyze tweets as positive, negative or neutral. These tweets are manually annotated with sentiment labels using Sentiment analysis API which uses Long Short Term Memory algorithm (LSTM).Naive Bayes Multinomial Text and the zeroR classifier is applied using Weka on the trained dataset. The accuracies of two classifiers are examined and compared. The result shows that Naive Bayes Multinomial classifiers show the better result when use training set test option is used with the balanced dataset.

Keywords— Social media, sentiment analysis, twitter, Weka, Naive Bayes Multinomial text classifier, zeroR classifier

I. INTRODUCTION

During crises, people face time-sensitive requirements like food, medical care [8].In this time social media plays a vital role in collecting information of the suffering people in particular areas. Social media has emerged as a popular medium for user-generated information and fast communication particularly during the crisis [7]. The large portion of this user-generated information consists of useless information but some of the available information can be used for detecting crisis [1]. Unfortunately, information broadcasted during the crisis is not possible for people to effectively find it and act on it without computational support [8]. For this reason, several researchers are working on sentiment analysis using social media so that these people may get informed and rescued as early as possible. Extracting people's opinions from textual data is called Sentiment analysis [2]. Sentiment analyses help to categorize tweets according to the concern and anxiety of people into positive, negative or neutral [8]. In this paper, we describe a methodology for collecting crisis related tweets and tagging them manually with their respective polarities. This paper explores the use of Twitter data in analyzing sentiments during hurricane Sandy 2012 between October 26, 2012 to October 30, 2012 [3]. The twitter data allows its user to share short messages called Tweets which are of maximum 160 characters [6]. So, the messages provided have good enough information for sentiment analysis. One example of sentiment analysis using social media is the American Red Cross Digital Operations Center, opened with Dell in March 2012 [1].

The rest of paper is outlined as follows. In the next section, we gave an overview of related work. Then it is followed by methodology section where we described how crisis related tweets have been collected and annotated to obtain training dataset. After that, we presented experimental results and its discussion by applying ZeroR and Naive Bayes Multinomial text classifiers using balanced and unbalanced datasets. Details regarding implementation are also discussed in this section. In the end, we concluded the paper along with suggesting some future work.

II. RELATED WORK

In [13], the authors conducted a comparison of sentiment polarity classification methods for twitter text. They performed the comparison using various classifiers. They used the number of manually annotated tweets as a training set. The polarities like positive, negative has been problematic since it does not cover the case of tweets with neutral polarity.

In [14], the authors described a system to detect political topics emerging in twitter. The main focus is on fast detection based on few tweets at an early stage of discussion. They also extended their work by detecting the polarity of topics marked by hashtags.

In [10], Sentiment analysis is performed using Decision trees and Support vector machines classifiers. Sentiment analysis using support vector machines classifier showed high accuracy when compared to Decision tress classifier. In Weka Decision tree classifier used was J48 and support vector machine classifier used was SMO.

In [11], authors used Naive Bayes and SVM classifier for sentiment analysis. Then they proposed a method using Naive Bayes and modified k means which was implemented using stacking classifier. Stacking classifier is used to cluster two or more than two classifiers together. Clustering Naive Bayes and SVM was found more accurate than Naive Bayes and Support vector machine techniques individually.

However, after a thorough investigation in the related scientific literature, we came up with the result that there has not been done the comparison of Naive Bayes multinomial classifier and zeroR classifier. Hence, we present the experiment using various test methods available in Weka. We also observed which classifier works best under certain test method.

III. METHODOLOGY

The following steps were involved in sentiment analysis using twitter data for disaster management.

A. Collecting tweets

The first step in the methodology was to collect a large set of crisis-related tweets. For this, we have downloaded existing crisis related collection from website www.crisislex.org. It contained tweets related to many crises that have occurred in past and we selected Sandy Hurricane tweets for sentiment analysis occurred in October 2012.

B. Annotation process and dataset format conversion

The dataset containing tweets is trained by annotating the tweets with the sentiments. The sentiments are categorized as positive, negative or neutral. As the dataset is large enough to be annotated manually it is sampled to 1600 for unbalanced dataset and 2400 for balanced dataset. The tweets are annotated using sentiment analysis API which uses Long Short Term Memory algorithm to set the polarity of tweets.

By default, the Weka uses the Attributes Relation File Format for data analysis. Other than this it accepts CSV format and Database using ODBC format. These two formats once loaded needs to be converted into the ARFF format for further processing of dataset. We downloaded our dataset in CSV format which is converted into ARFF format by following few steps using Weka. The sample dataset in ARFF format is shown in Table I.

TABLE I
SAMPLE DATASET IN ARFF FORMAT

```
@relation sentiment
@attribute tweet string
@attribute sentiment {positive,negative,neutral}
@data
"I've got enough candles to supply a Mexican family",positive
"Sandy be soooo mad that she be shattering our doors and shiet
HurricaneSandy",negative
"thankfully Hurricane Waugh played it cool and waited this one out",positive
"you never got that magnificent case of Burgundy I sent you to thank you for your
tweets?",negative
"I'm at Mad River Bar",neutral
"Cory_Kennedy arrives to the rescue sporting some spelunking equipment
#sandy",positive
"And that's it until the spring",positive
```

The first line in ARFF file starts with @relation written in the header section, followed by the dataset name i.e sentiment. Next line contains @attribute which is followed by attribute names i.e tweet and sentiment along with respective datatypes. @data indicates the starting of data.

C. Data preprocessing

We took several preprocessing techniques available in Weka to clean the tweets before applying classifier.

1) *Stopwords handler:* It is the technique which eliminates frequent usage of words which are meaningless and useless for text classification. We used MultiStopwords option for our experiment.

2) *Stemming:* stemming removes the suffix of the word by reducing it into its root form. For example, read, reader, readers, reading all reduce to the root word read. We used lovin Stemmer in our experiment.

3) *Tokenizer:* This splits the document into words/terms, constructing a word vector known as bag-of-words. We used NGramtokenizer in our experiment.

D. Classifier

Weka is a software that is provided with the number of machine learning classifiers. Well, known classifiers, Naive Bayes Multinomial Text and zeroR classifiers available in Weka are used to perform experiments using unbalanced and balanced dataset. The accuracies are compared using the different test options available in Weka tool.

IV. EXPERIMENTAL RESULT & DISCUSSION

As there is the limit in how many tweets can be tagged manually in a reasonable time, we trained 2400 tweets for a balanced dataset and 1600 for an unbalanced dataset. For a balanced set, we used 800 tweets for each category i.e positive, negative or neutral. First, we used unbalanced dataset for sentiment analysis and applied classifiers after preprocessing. The Fig.1 shows the attributes present in dataset i.e tweet and sentiment. It also visualizes number of positive, negative and neutral tweets using bar graph.

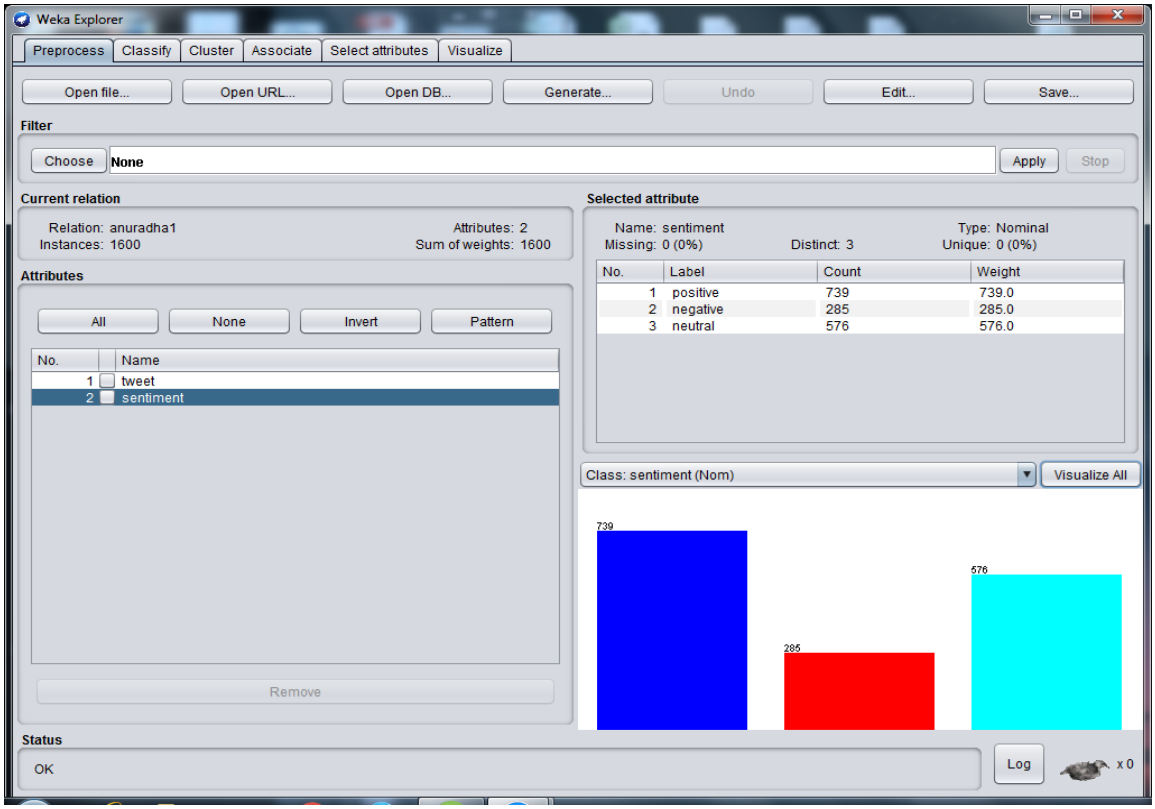

Fig.1 Unbalanced dataset

We performed experiments using different test options, using Naive bayes multinomial classifier and zeroR classifier to estimate the accuracy of the classifiers. The accuracy obtained using unbalanced dataset is shown in Table II.

TABLE II
ACCURACY RESULTS USING UNBALANCED DATASET

	Use training set	**Cross validation(10 fold)**	**Percentage split**		
zeroR classifier	46%	46%	70 45%	80 46%	90 48%
Naive Bayes multinomial text classifier	75%	52%	55%	56%	59%

The results show that Naive Bayes multinomial text classifier performed better than zeroR classifier. If we compare single classifier using different test options, zeroR gave better result using percentage split of 90% .The Naive Bayes multinomial text classifier performed well using use training set test option and gave a good accuracy.

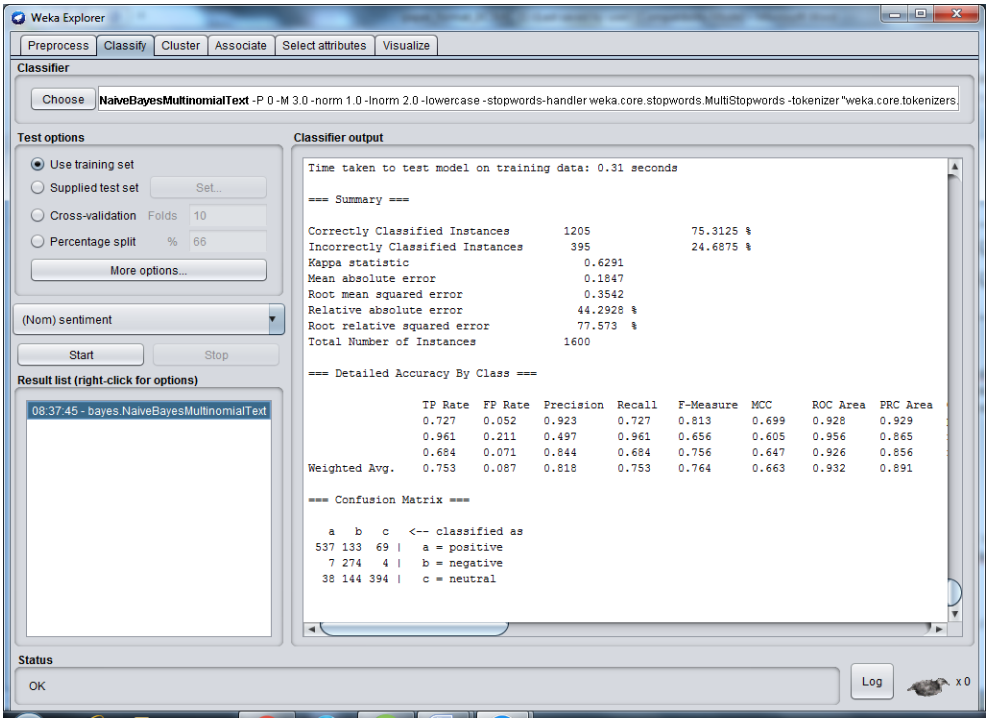

Fig.2 Weka screenshot for Naive Bayes Multinomial Text classifier using unbalanced dataset

After estimating the performance of classifiers using unbalanced dataset, we tried to improve the accuracy by using balanced dataset. So, we annotated equal number of positive, negative and neutral tweets as shown in Fig.3.

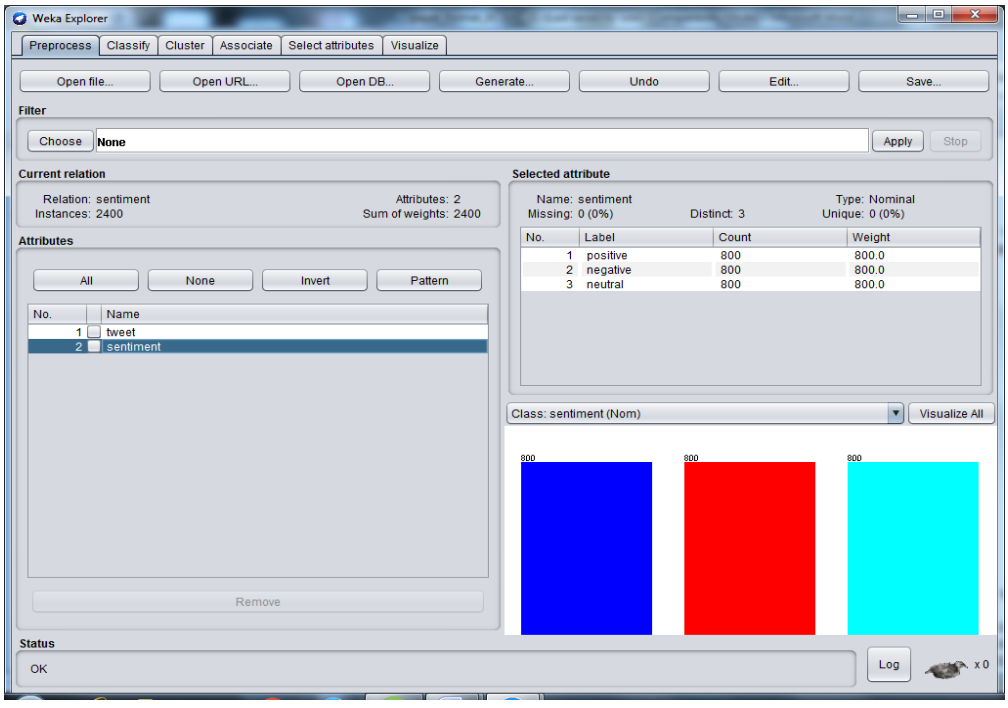

Fig.3 Balanced dataset

When we used balanced dataset, the performance of zeroR degraded whereas the Naive Bayes showed better results. Naive Bayes Multinomial text classifier showed improvement in accuracy. Naive bayes showed best results when use training set option is used. The accuracy obtained using unbalanced dataset is shown in TableIII.

TABLE III
ACCURACY RESULTS USING BALANCED DATASET

	Use training set	**Cross validation(10 fold)**	**Percentage split**		
zeroR classifier	33%	33%	70 30%	80 30%	90 28%
Naive Bayes multinomial text classifier	84%	58%	57%	57%	58%

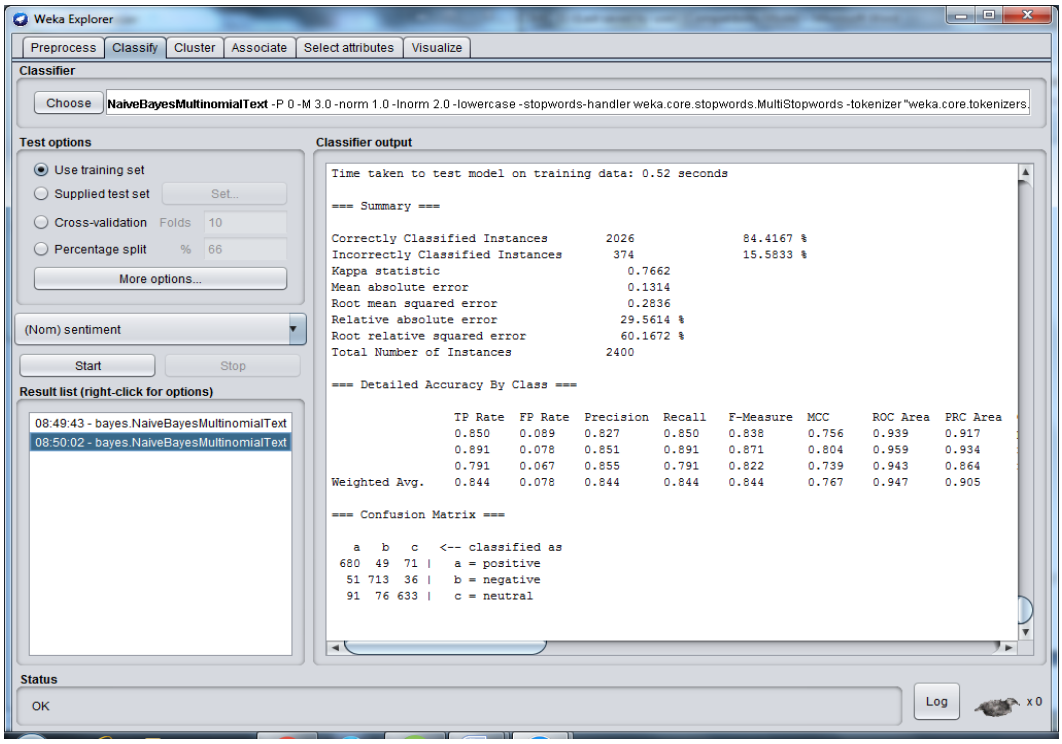

Fig.4 Weka screenshot for Naive Bayes Multinomial Text classifier using balanced dataset

It was observed in Fig.4 that dataset used should be balanced, suitable preprocessing techniques must be applied before applying machine learning classifier to obtain high accuracy. Although the results are promising it can be questioned that whether the obtained accuracy is good enough to be used in real world sentiment analysis for crisis management. We believe that results are good enough to be used on aggregate level.

V. CONCLUSION AND FUTURE WORK

We have described a methodology for collecting hurricane Sandy disaster tweets and tagging them manually according to their respective polarities. The resulting dataset is used as a training set and is converted into Weka compatible format. After applying classifiers it was observed that Naïve Bayes Multinomial Text classifier with 84% accuracy performs better than zeroR with 75% accuracy using use training set option to test data. It was also observed that using appropriate preprocessing technique also improves the accuracy.

However, it will be good to investigate the available preprocessing techniques in order to get optimal results. More training data would also improve the accuracy but high cost in terms of manpower should also be in limit. It is also expected that better performance should be achieved using unbalanced dataset this means their should be no requirement to balance the dataset.

REFERENCES

[1] 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

[2] Abdullah Alfarrarheh, Sumeet Agrawal, Seon Ho Kim, Cyrus Shahabi, "Geo-spatial Multimedia Sentiment Analysis in Disasters", International Conference on Data Science and Advanced Analysis (DSAA 2017), IEEE Oct. 2017

[3] Han Dong, Milton Halem and Shujia Zhou,"Social Media Data Analytics to Hurricane Sandy" International Conference on Social Computing,IEEE Sept.2013

[4] Rafet Duriqi, Vigan Raca, Betim Cico, **"**Comparative Analysis of Classification Algorithms on Three Different Datasets using WEKA"5th Mediterranean Conference on Embedded Computing (MECO)",IEEE June 2016

[5] 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

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

[7] Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark Cameron, Bella Robinson, Robert Power "Using Social Media to Enhance Emergency Situation Awareness Extended Abstract", Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

[8] Sudha Verma, Sarah Vieweg, William J.Corvey, Leysia Palen, James H.Martin, Martha Palmer, Aaron Schram and Kenneth M.Anderson"Natural Language Processing to the Rescue? Extracting "Situtational Awareness" Tweets During Mass Emergency", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM) 2011

[9] Venkata K.Neppalli, Cornelia Caragea, Anna Squicciarini, Andrea Tapia, Sam Stehle "Sentiment analysis during Hurricane Sandy in emergency response", International Journal of Disaster Risk Reduction 21 (2017) pg.213–222

[10] Umadevi V,"Sentiment analysis using Weka",International journal of Engineering Trends and Technology(IJETT)-Volume 18 Number 4-Dec 2014 pg:181-183

[11] Preety, Sunny Dahiya,"Sentiment Analysis Using SVM and Naive Bayes algorithm", International Journal of Computer Science and Mobile Computing, Vol.4 Issue.9, September- 2015, pg. 212-219

[12] Jaspreet Singh, Gurvinder Singh and Rajinder Singh, "Optimization of sentiment analysis using machine learning classifiers",Human-centric Computing and information Sciences 2017,SpringerOpenJournal

[13] 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

[14] 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.

