Support Vector Machine and Naïve Bayes comparison of Sentiments on Terrorism

Muhammad Umer Haroon

ABSTRACT

Text Analysis has become a major area of research. In order to be aware of what people think and how they feel after terrorism attacks, there needs to be some mechanism. We aim to propose a solution in this regard to learn about people's sentiments in detail on terrorism incidents in Pakistan using text analysis. In this research support vector machines and naïve Bayes algorithms are compared in finding out the sentiments from data set of opinions express on terrorism activities in Pakistan.

General Terms

Support Vector Machines, Naïve Bayes, Twitter.

Keywords

Sentiments, Text analysis, terrorism incidents.

1. INTRODUCTION

In any language a combination of words that portray a meaning are called as text. Text analysis has become a major area of research as data is increasing at a rapid pace. Every social media and social tools are encouraging people to share their views about issues, products and services. This enormous interaction generates gigantic amounts of data on web. Data generated is huge but is only demanding more storage resources if not brought into use. The analysis of text data is called text analysis. This analysis of text data is fruitful in many areas. Text analysis of data has been proving beneficial in many areas including business, military, medicine and for governments. The field of text analysis is very broad and encompasses many sub fields. Text analysis includes topic modeling, text summarization, opinion mining and sentiment analysis.

The focus area in this research is sentiment analysis. Traditionally sentiments mean feelings. Text is very rich source of information having many useful patterns sentiments are one of them. In sentiment analysis a text is analyzed for the feelings expressed in it usually positive and negative. There are multiple methods for performing sentiment analysis with each using different logics and working.

Sentiment analysis is performed by machine learning, linguistics and natural language processing.

2. PRIOR WORK

The text analysis about sentiments of people of Pakistan has not been done before and specifically in regard to tragic terrorism incidents. Internationally studies have been done analyzing the feeling of people and have shown how important it is to tell about the people after disasters [1]. Though data analyzed is about an accidental disaster not terrorism activity. The current work on getting sentiments using text analysis is only limited to the polarity of opinions and not the granularity [2]. The study by Andrea et al. [3] shows that social media information can be effectively used by governments to help them understand the people opinions in emergency situations. Research studies about terrorism have significantly increased after 9/11. Nizamani et al. [4] has done notable work using text analysis to find out the incident type of terrorism activity. Text mining is used for military and security purposes [5]. The people feelings are very important to be analyzed after disasters in order to avoid anarchy and mishaps. Woo et al. [1] study is thought about a specific disaster and shows how people have reacted to that incident.

There are huge reactions from people suffering from the trauma of terrorism activities. The work of Schlenger et al. [6] shows that general public other than directly affected also suffer from incidents of terrorism.

A vital research by Neviarouskaya et al. [7] specifies a tool for more granular analysis of text to find the emotions and not only positive, negative and neutral polarity.

3. PROCEDURE

For performing sentiment analysis data chosen was tweets from twitter. Twitter has become an important platform for discussing issues and topics in real time online. Terrorism activities are increasing in Pakistan day by day and have been consistently occurring since Afghan war [8]. Terrorism incidents are affecting everything in country: the economics, lifestyle and attitude. The change in people attitude after every such incident is mixed of emotions for instance grief, revenge, hope and stress. There is a bulk of text data generated after every such incident and people talk about it especially on social media and online web. Journalists and authorities give their own analysis and publish their opinions and so does the public. With every day such events are increasing in Pakistan and affecting the life of citizens in one way or another. Tweets were fetched from twitter about such incidents using hashtags.

There are multiple ways of performing sentiment analysis and machine learning is one of the known approaches. In this research Support Vector Machines and Naïve Bayes are selected for performing sentiment analysis. Support vector machine is bi classifier algorithm. In Naïve Bayes classifier the order of words does not makes difference to sentiment predicted from the text, it takes words on a count as whole. The equation for Naïve Bayes is

$$P = \frac{P(C)}{V_C^n} \cdot \prod_i^n count(d_i, C)$$

4. EXPERIMENT

Data was collected from twitter using keywords and hashtags about terrorism incidents from year 2016. Around 30,000 tweets were collected from random incidents where people expressed different sentiments online. Majorly all tweets collected were about bomb blast incidents affecting a lot of people resulting in casualty. Some of the incidents include Allama Iqbal Town bomb blast on March 27, 2016 and on August 8,2016 Civil Hospital Quetta blast. Example of collected tweets shown in Table 1.

Table 1 Sample of Tweets from collected data

Polarity	Tweet
negative	Every lawyer that has ever given me a lift home is dead, except for one, Naveed Qambrani, he is critical and was airlifted to Karachi...
positive	Amazing scenes as people queue to donate blood in #Lahore
neutral	My deepest & heartfelt condolences to all.Attacks by the unjust on lawyers should only strengthen our resolve for justice

Three polarities were set for text positive, negative and neutral. Positive polarity means the text conveys a good feeling, negative polarity means the text is expressing sadness, anger and feelings of dislike. In neutral the emotions expressed in text just cancel the effect of each other like having words from both positive and negative feelings. Neutral text also may not contain any of the words inclined towards a specific sentiment.

The data was gathered by use of search API of twitter and manually tweets were collected whenever such incident happened.

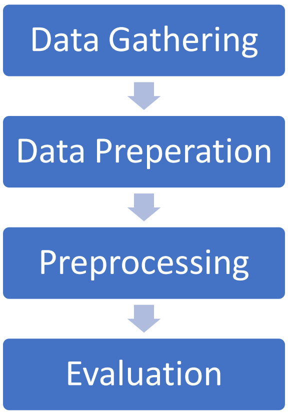

Fig.1 Methodology Flow

The gathered data tweets were reformatted and only text was obtained removing unwanted information as publisher of tweet and locations. The prepared text data was set for preprocessing. In preprocessing phase, the non-textual characters which do not add meaning to text were removed. The last phase was to perform evaluation on the refined data to obtain sentiment analysis results and the process is shown in Figure 1.

The collected tweets were analyzed by using Support Vector Machine and Naïve Bayes algorithm in WEKA. WEKA is a software which has a combination of machine learning algorithms.

Since our data set is preprocessed it is easy to obtain polarities from it. The actual data obtained was unclassified and was marked with polarities manually by 15 persons and was a hard task. Average of collectively marked data was taken, and the standard threshold was set for experiments. The values set for all data were 19800 tweets as negative which is 66% ,6300 positive which is 21% and remaining 3900 as neutral which is 13% of whole data. The 30,000 tweets had the polarities ratio as 77% of them as negative 14% positive and 9% neutral on single fold results for support vector machine. The data was again tested for two mean folds and results were a bit different this time. All values can be seen in table 2 for 1-fold and 2 folds. The results differentiated in positive and neutral polarity largely and small difference for negative polarity.

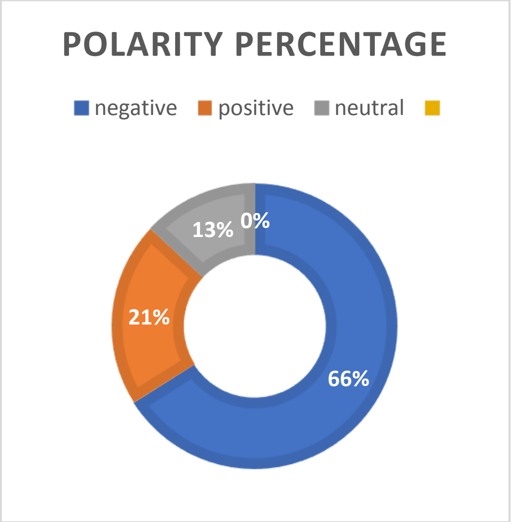

Fig.2 Polarity ratio of all data set as standard

In case of Naïve Bayes, the obtained results can be seen in table 3. For Naïve Bayes the results for positive polarity in 2 folds is very accurately classified.

For validating the obtained results various parameters can be set and many methods can be adopted. In this research three criteria were selected mostly used for machine learning research precision, recall and accuracy. Precision means properly selected data from whole data set classified as correct. Recall means sum of both correctly selected data and not correctly selected which should have been classified from whole dataset. The precision, recall of results calculated is summarized in table 4 for SVM. The table 5 shows precision and recall values for NB for 1 and 2 folds.

Table 2 Per Fold Results for SVM

	negative	positive	Neutral
1-fold	77	14	9
2-folds	73	11	16

Table 3 Per Fold Result for NB

	negative	positive	Neutral
1-fold	63	17	20
2-folds	71	21	8

Table 4 Precision and Recall Values for SVM

polarity	1-fold Precision	2-fold Precision	1-fold Recall	2-fold Recall
negative	0.857	0.904	1	1
positive	1	1	0.66	0.52
neutral	1	0.81	0.69	1

Table 5 Precision and Recall Values for NB

polarity	1-fold Precision	2-fold Precision	1-fold Recall	2-fold Recall
negative	1	0.92	0.95	1
positive	1	1	0.8	1
neutral	0.65	1	1	0.61

5. CONCLUSION

The results obtained from tweets and sentiments extracted via Support vector machine and Naïve Bayes by using validation criteria of precision and recall express that both algorithms work well with limitations. Also, there is inverse relation between precision and recall increasing or decreasing one values effects other. The data set nature can also affect the results obtained but in this case, it is evident that Naïve Bayes has more accurate values than Support Vector Machine. The difference is marginal and vary with multiple factors.

This research can be further improved by using more folds for obtaining accurate results and use of annotated data set can cause a varying result. The overall sentiments expressed over terrorism incidents by users on twitter is highly negative expressing sad and condemning behavior. This sort of data can be usefully used for treatment of trauma patients.

6. REFERENCES

[1] H. e. a. Woo, "Public trauma after the Sewol Ferry disaster: The role of social media in understanding the public mood.," International journal of environmental research and public health, vol. 12, no. 9, pp. 10974-10983, 2015.

[2] A. P. P. Pak, "Twitter as a corpus for sentiment analysis and opinion mining.," vol. 10, 2010.

[3] A. L. a. F. E. A. a. S. S. D. a. Y. S. a. L. L. T. ,. D. J. ,. N. A. ,. L. Kavanaugh, "Social media use by government: From the routine to the critical," Government Information Quarterly, vol. 29, pp. 480-491, 2012.

[4] N. ,. S. Memon, "Analyzing news summaries for identification of terrorism incident type," Educational Research International (erint.), vol. 3, no. 4, pp. 81-88, 2014.

[5] B. ,. S. L. Z. Haarmann, "Applied Text Mining for Military Intelligence Necessities," in Proceedings of the 6th Future Security Conference, Berlin, 2011.

[6] W. E. a. C. J. M. ,. E. L. ,. B. K. ,. K. M. ,. D. a. T. L. ,. D. J. M. ,. F. J. A. ,. R. A. Schlenger, "Psychological reactions to terrorist attacks: findings from the National Study of Americans' Reactions to September 11," American Medical Association, vol. 288, no. 5, pp. 581-588, 2002.

[7] A. ,. A. M. a. P. H. ,. I. M. Neviarouskaya, "Intelligent interface for textual attitude analysis," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, no. 3, p. 48, 2014.

[8] "worst-terrorist-attacks-pakistans-military-forces," yumtoyikes, [Online]. Available: http://yumtoyikes.com/2015/10/05/worst-terrorist-attacks-pakistans-military-forces-since-911-part/. [Accessed 10 2 2016].