Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-5, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Terror Detection Using Text Mining

Prof. Shriram Kulkarni, Sayali Baban Pawar & Ganesh Satpute
Department of Information Technology and Engineering
Sinhgad Technical Education Society Savitribai Phule Pune University, Pune, Maharashtra, India

Abstract: *Web or World Wide Web is storage for huge amount of information. The Users of the web can access information stored on web via computers. The main reason behind rapid growth of web is social web culture. Every day the number of web users is increasing, hence parallel growth in Cybercrime. Web allows us discussions and sharing of our knowledge freely, but it also has a negative side which can be harmful for society. Sharing or discussing information related to terrorism is one of the most important problems faced by many countries today. Terrorists are using web as a medium to spread their activities and encourage young generation to join them. To find a solution to this problem this paper presents the implementation of a system which detects terror activities on web by using text mining.*

Keywords: *Text Mining, Terror Detection, Web.*

1. Introduction

The Terror Detection system finds all of the violent or malicious or terror related comments/discussion on web page. This is a very challenging task as the web page contains not only comments but other stuff such as advertises, header, footer, images and web addresses etc. The very first step is to extract all of the data on a web page. Extracted data is not pure; it contains lots of unwanted data which needs to be removed from extracted data. To get actual text data pre-processing is done on original data. The end result of pre-processing step is in the form we need, pure text data.

Even this pure text data contains lots of unwanted words which has negligible importance in sentence. Such words are often called as 'Stop Words'. A list of 'Stop Words' is maintained as is given to system to delete those words from comments. This removal of high frequency words such as 'the', 'a', 'is', 'are' etc reduces the size of the actual data and makes processing easy. Now the classification of remaining actual text data is done. The comment gets classified as 'violent' or 'non-violent'.

Many different algorithms are available to find the important words in a text comment. TF-IDF is one of them and it is simple to use. We will discuss about it later in this paper. Also to find similarity between different samples, similarity measure algorithms are used. These two are the most important factors of terror detection system.

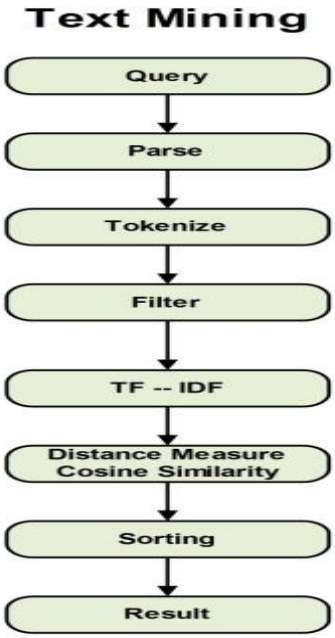

Fig. 1 Text Mining Steps

2. Extraction

Automatically getting structured information from semi or unstructured information is known as' Extraction'.

2.1 OAuth

With the growth of web, dependability on distributed systems and cloud computing is increasing. The problem in a third-party application is, to access data on other sites, they require username and password. This may lead to exposing user password to someone else.

A solution to this problem is OAuth. It provides a method for users to grant third-party access to their resources without sharing their passwords. It also provides a way to grant limited access. It is commonly used as a way to log into third-party websites using Microsoft, Google, Facebook, Twitter,

One Network etc. accounts without exposing their password. OAuth essentially allows access tokens are issued by an authorization server to third-party clients, when resource owner allows it to. The third party then uses the access token to access the protected resources hosted by the resource server.

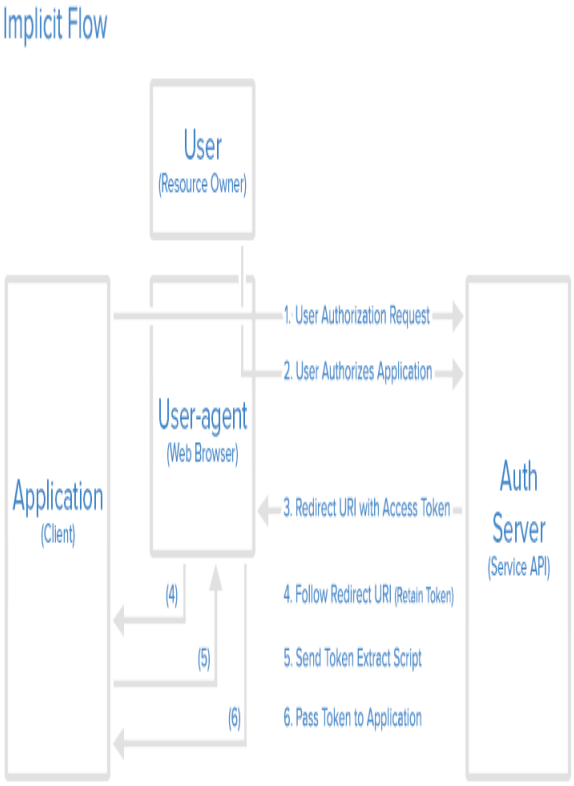

Fig.2. Implicit Flow of OAuth system

3. TF_IDF

IT-IDF is **T**erm **F**requency-**I**nverse **D**ocument **F**requency. TF-IDF weight is used in text mining to judge importance of a word for a document in a collection. Mathematically this can be expressed as follows

$$wd = fw, d * \log (|D|/fw, D)$$

Where, *D*- is given document collections,
w - a word,
d - an individual document such that $d \epsilon D$
fw, d- number of times w appears in d,
|*D*|- *size* of the corpus,
fw, D- number of documents in which w appears in D

Simply,

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF (t) = log_e(Total number of documents / Number of documents with term t in it).

4. Cosine Similarity

Cosine similarity is used to find the similarity between two documents or queries or a document and a query. Cosine Similarity generates a metric that gives how two documents are related by looking at the angle.

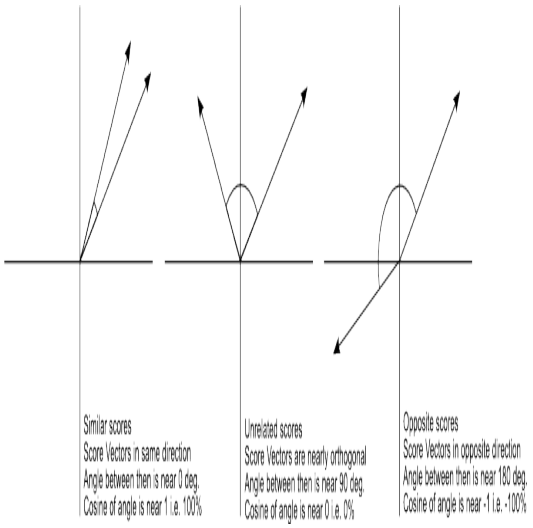

Fig.3. the Cosine Similarity values for different documents

An angle of degree 0 means documents are similar, where as angle of degree 90 represents two different documents. The following term gives the mathematical formula for cosine similarity.

$$\frac{|X \cap Y|}{|X|^{1/2} . |Y|^{1/2}}$$

Fig.4. Mathematical formula for cosine similarity

Where, X- any of the documents in a group
Y- Corresponding query

5. Dice coefficient

Dice's coefficient is similarity measure used as alternative for similarity cosine. The results computed by both of these algorithms are different.

Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings. The following fig. gives mathematical formula to compute Dice's coefficient.[[2]]

$$\frac{|X \cap Y|}{|X|^{1/2} . |Y|^{1/2}}$$

Fig.5. . Mathematical formula for Dice's coefficient

Where, X- any of the documents in a group
Y- Corresponding query

6. Jaccard

Jaccard is another similarity measure and it is computed as the number of shared terms over the number of all unique terms in both strings[2]

$$\frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

Fig.5. Mathematical formula for Jaccard coefficient

Where, X- any of the documents in a group
Y- Corresponding query

7. Experiment work

In our experiment we downloaded comments from a social networking site, stored them in database. The lists of violent, non-violent and stop words are given to the system. E.g. violent words such as 'Terrorist', 'bad' etc.

When a comment is downloaded it's TF-IDF score is calculated and based on similarity measures the comment is identified as either violent or non-violent. E.g. The comment is

"Terrorists are bad and I'm a terrorist."

The system will calculate it's TF score as follows

7.1 TF score-

terrorists:::2.0
are:::2.0
bad:::0.41503749927884376
and:::1.0
m:::2.0
a:::2.0
terrorist:::2.0

7.2 IDF score-

movie::2.0
very::1.3333333333333333
dirty::1.3333333333333333
happy::0.6666666666666666
hello::0.3333333333333333
am::0.3333333333333333
adult::0.3333333333333333
scene::0.3333333333333333
bad::0.1383458330929479
is::0.1383458330929479
movie::2.0
very::1.3333333333333333
dirty::1.3333333333333333
happy::0.6666666666666666
hello::0.3333333333333333
am::0.3333333333333333
adult::0.3333333333333333
scene::0.3333333333333333
bad::0.1383458330929479
is::0.1383458330929479

7.3 Similarity measures result

COSINE SIMILARITY:::: Document is violent
DICE::::::::Document is violent
JACCARD>>>>>>>>Document is violent

8. Results

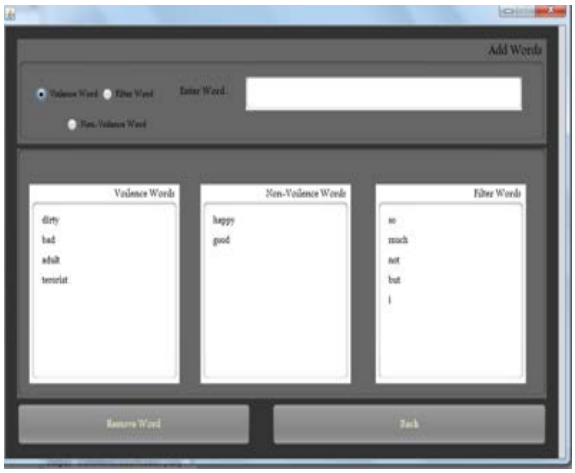

Fig.6 Add words to lists of violent, non-violent and filter words

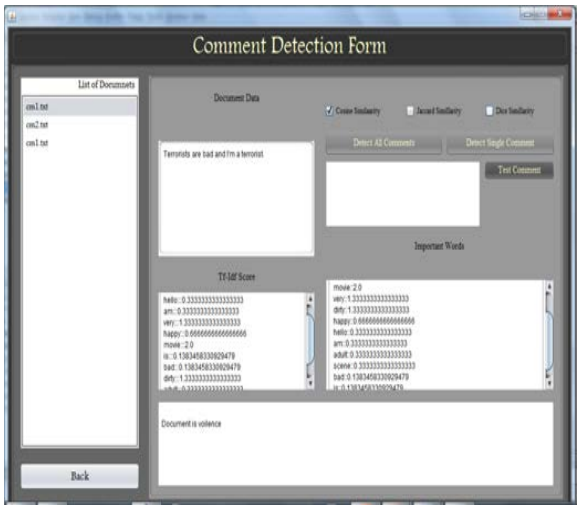

Fig.7 Calculate IF-IDF score and classify comment using similarity measures (violent comment)

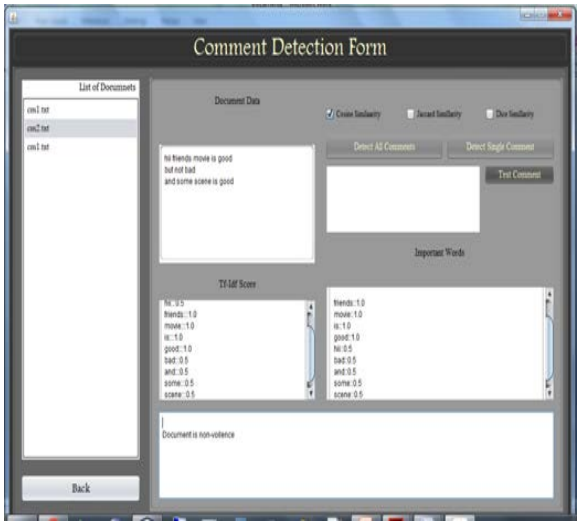

Fig.7 Calculate IF-IDF score and classify comment using similarity measures (Non-violent comment)

9. Conclusion

We have implemented a system which detects the terror related comments on social networking sites and will help for national security. We can improve its efficiency by combining three similarity measures. The system can also track comment, when a particular user is tying. System will classify comment while user is typing and will block user from posting it, if it is violent.

10. References

1. Vikas Thada, Dr Vivek Jaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm" International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013 202 SSN: 2319-1058
2. Wael H. Gomaa, Aly A. Fahmy," A Survey of Text Similarity Approaches" International Journal of Computer Applications (0975 – 8887) Volume 68– No.13, April 2013
3. D. Michelangelo, C.France, L.Steve, C. Lee, G. Macro, "Focused crawling using context graphs" , Proceedings of the 26th international conference on very large database , pp. 527-534, 2000
4. N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.
5. E. Martin Ester, G.Matthias, K. Hans-Peter Kriegel, " Focused Web Crawling, " A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies " Proceedings of the 27th International Conference on Very Large Database, pp.633-637,2001