Detecting Terrorist Activities using Sentiment Analysis In a Distributed System

Manas Kocharekar, Umesh Jadhav
Dept. Of Information Technology
Bharati Vidyapeeth College of Engineering, Navi Mumbai,India.
kocharekarmanas@gmail.com; ujadhav25@gmail.com

Abstract

Terrorism is a growing problem in today's world. Organizations like ISIS, Taliban, Lashkar-E-Taiba and many nameless others are spreading their works at an accelerated speed. These operations are now not only limited to arms dealing and bombings. The terrorists have mastered the new powerful tool called "the social network". Terrorists have taken it to the social network to do their brainwashing and organizational expansion because it is the easiest way to reach the masses and impact people's opinions. Therefore, it is crucial to monitor this form of activities to minimize the impact of them on naive youngsters who blindly follow the provoking leaders.

I. INTRODUCTION

Sentiment Analysis is a technique that uses NLP, statistics and Machine Learning methods to extract,identify, or otherwise characterize the sentiment of a text unit. It's also defined as the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. It is a branch of artificial intelligence, not to be confused with Sentiment Mining which focuses on calculating` the general opinion of a particular group of people, extracting predefined keywords to check their frequency & perform statistical algorithms. The sentiment Analysis on the other hand works deeper than that. It actually tries to find out the interrelation between words of text or the corpus and thereby decides its sentiment value. When teamed up with Machine Learning, it can train itself to identify meanings from the text along with newly discovered words & patterns. Thus, the results of sentiment analysis get better with the amount of text it processes.

For Machine Learning tools to process and identify the emotions from the Social Posts as mentioned in the abstract of the report, we also need to extract the text content from the social media. Now, if we consider the amount of information that is shared every minute on different platforms, we will know that traditional RDBMS are not going to be able to process that kind of data. This is where Big Data tools come in. Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Which means that instead of having an expensive server to store and process large data, Apache Hadoop makes use of multiple cheaper computer systems and distributes the big task to each of them in a divided fashion which makes it fast, efficient and most importantly Fault Tolerant.

II. ARCHITECTURE DIAGRAM

Social Networks: These are the various sources that will provide the data feed as an input to the application. It will include Facebook, Twitter, Instagram etc. These social networks provide their APIs so that data can be fetched easily to the application servers.

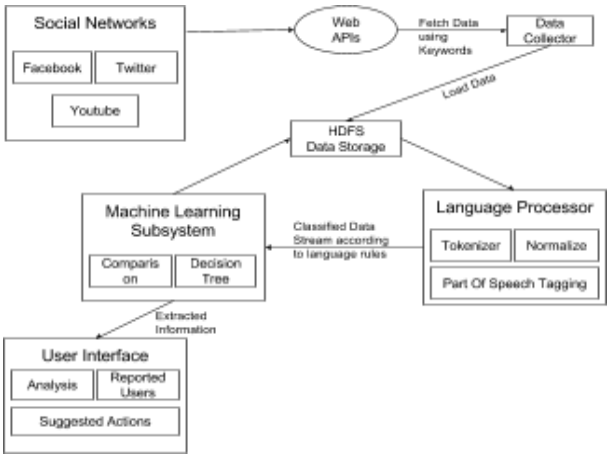

Fig.1 Architecture of the System

Web APIs: Web APIs are the applications that provide means for exchange of data between multiple web servers. They can be programmed to pass on requested information over the internet. Here the APIs provided by social networks will let us collect selective data from millions of posts on the social networks.

Data Collector: This unit will be responsible for communicating with the APIs & pass on the collected selective data into the HDFS.

HDFS Storage: Apache Hadoop provides HDFS (Hadoop Distributed File System) that stores unstructured data efficiently with minimum write cost & time.

Language Processor: It covers aspect of Artificial Intelligence that deals with extracting the meaning from the text using various techniques such as tokenization, normalization, common words removal, part of speech tagging etc.

Machine Learning Subsystem: Once the text input is processed our logic will try to teach the machines to determine the sentimental value of that text & in turn produce the flagged content.

User Interface: Here all the statistics are displayed on the web based front end UI, from where the Administrator can produce reports & call for action accordingly.

III. EXISTING SYSTEMS

There aren't many tools implementing the AI to perform sentiment analysis for security or law enforcement as the line between a normal post and a sensitive post is very fine and most times it's almost impossible to predict the actual intent of that post. Following are the two systems that have been at top of this vertical.

Parameters	**Geofeedia**	**SMC by IBM**
Social Networks Covered	Twitter, Instagram, Facebook, flickr, Youtube	Twitter, Facebook, Google, Classifieds website
Area of Focus	Geographical Location	configured "surveillance targets"
Pricing	High	Too High
Usability Complexity	High	High

IV. Proposed System

The proposed system is a web application that will extract data from the social media and feed it to the Apache Hadoop for processing it further, which will be running on a cluster of commodity hardware. Once the data is processed it will go through the Sentiment Analysis logic which will include Natural Language Processing (NLP) and Machine Learning tools written in Python. The final results or the alarms raised by the system will be shown in the front end of the application which will be a Web Interface again based on Python Framework.

The system is to be used by an authorised government official only as it can prove a dangerous tools in wrong hands given the power it will have for processing and understanding the sentiments in social post. The authorization will be enforced to access the system. The system will generate the reports on daily frequency of flagged content, its source, GeoLocation, timestamp and severeness of that content.

It will show the accounts frequenting in the flagged content, so that the action could be taken against them accordingly.

The system covers only text based posts. It cannot process Pictures or Video content as of now. The system will also not be able to process the non-English languages on its completion, but, it can certainly be part of its future scope.

V. CONCLUSION

Although there is some work going on in sentiment analysis for business marketing, not much work is being done in the area of crime control or fight against terrorism. The content shared on social media by extremist organizations is playing a major role in their recruitment of youth. We should accept fact that future fight against terrorism will be fought on the internet. We have to come up with a technology to monitor the social networks and keep them clear of such extremist content so that real power of social networks can be harnessed by connecting more people online. We hope that our system, on its completion is able to extract the real meanings behind the social posts to predict their sentiment value. It will also be able to train itself to understand the other offensive words that are not included in its list & add them automatically to enhance the overall system vocabulary.

REFERENCES

[1] “Apache Hadoop Goes Realtime at Facebook” 2010 [Borthakur, Muthukkaruppan]

[2] “Large Scale Sentiment Analysis for News and Blogs” [Godbole,Srinivasaiah,Skiena]

[3] “Sentiment Analysis With Global Topic and Local Dependency” 2011 [Li,Huang,Zhu]

[4] “Thumbs up? Sentiment Classification using Machine Learning Techniques” [Bo Pang, Lillian Lee, Shivakumar Vaithyanathan]

[5] “Real Time Sentiment Analysis Of Twitter Data Using Hadoop” 2014 [Sunil Mane, Yashwant Sawant, Vaibhav Shinde]