# Twitter Based Event Detection System

Prof. Digambar Jadhav ,Ankit Kumar , Mimansa Singh , Rajan Yadav , Shubham Ghare.

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY , PUNE .

***Abstract— In this paper, a scheme is developed which can detect what happens in real world in real time only by analyzing tweets as Big Data and let a user know the event. The proposed scheme is mainly divided into three parts which are :***
***1. Tweet Preprocessing (TP) : It uses Twitter Streaming API and the Morphological Analysis Engine for the purpose.2. Importance Degree Calculating (IDC) : It is designed for getting Extended Hybrid TF-IDF values of the nouns and an average of the values. 3. Remarkable Words Detecting (RMD) : is designed for detecting remarkable words based on currently gotten Extended Hybrid TF-IDF values and some averages of currently and previously gotten them. User can click on any of the remarkable words shown to see the tweets regarding it.The project aims to demonstrate how Twitter Data Analysis can be used to give a clear picture about a particular event, so that the user can get an overall idea about events happening currently.The contribution of our project towards the technology is that we are demonstrating how the power of Data Analytics can be used for the benets of the general public.***

***Keywords— TP (Tweet Processing), IDC (Importance Degree Calculating ), RMD(Remarkable Words Detecting), TF-IDF.***

## 1. Introduction

The project aims to demonstrate how Twitter Data Analysis can be used to give a clear picture about a particular event, so that the user can get an overall idea about events happening currently.The contribution of our project towards the technology is that we are demonstrating how the power of Data Analytics can be used for the benefits of the general public.

Twitter has evolved to become a source of varied kind of information. This is due to nature of twitter on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express themselves. Thus, Twitter is the best microblogging platform used for event detection program.

## 2. Motivation

Because of a free format of messages and an easy accessibility of microblogging platforms, Internet users tend to shift from traditional communication tools to microblogging services. Traditional media can be in some or form inaccurate. The events in newspapers are published at a delay of minimum 20-24 hours. Also, the news reported by the Television media takes minimum 3-4 hours to reach the people. But here, using microblogging, the opinions are directly expressed from the people as soon as they wish to blog. As more and more users post about products and services they use, or express their political and religious views, or express and directly report a current event/incident happening around them, microblogging websites become valuable sources of peoples opinions and sentiments.With its over 240 million users tweeting out more than 500 million messages daily, Twitter is shaping public opinion like never before. Twitter has evolved to become a source of varied kind of information. This is due to nature of twitter on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express themselves. Thus, Twitter is the best microblogging platform that can be used for event detection.

- Hence, we need a tool to analyze these opinions, reviews, sentiments and frame all related tweets together to detect an event/discussion that will help the user to know about general opinion of the mass.
- The main contribution of our project towards society is that it is equipping it with a power

tool of analyzing the social media about events and getting an overall analysis about that topic.

• The contribution of our project towards the technology is that we are demonstrating how the power of Data Analytics can be used for the benefits of the general public.

## 3. Objective

Using proposed system, the following objectives are to be fulfilled :

1) To build a system which does not requires set of predefined keywords. Hence develop a system that detects and creates a set of real time trending keywords on its own.
2) To overcome problem of quantifying importance of words accurately.
3) Evaluating the quantified values dynamically.

## 4. Literature Survey

**"Detecting Real Time Events using Tweets Koichi Sato, Junbo Wang, Zixue Cheng"** According to author, Big Data has been one of main topics in the field of computer science. Additionally, demand for observations of the real world in real time has increased to provide services or information to people accordingly. For example, when disaster occurs, government can appropriately respond to the disaster if the situations in the disaster-stricken areas are realtimely grasped. Although there are many kinds of blog services and they are functioning as one of Big Data source, Twitter is considered as the most active Big Data source.

**"TEDAS: a Twitter based Event Detection and Analysis System Rui Li, Kin Hou Lei, Ravi Khadiwala, Kevin Chen- Chuan Cheng"** Witnessing the emergence of Twitter, we propose a Twitter-based Event Detection and Analysis System (TEDAS), which helps to (1) detect new events, to (2) analyze the spatial and temporal pattern of an event, and to (3) identify importance of events. In this demonstration, they show the overall system architecture, explain in detail the implementation of the components that crawl, classify, and rank tweets and extract locations from tweets, and present some interesting results of system.

**"A Survey of Techniques for Event Detection in Twitter - Farzindar A Tefeh and Wael Khreich."** According to author, Twitter is among the fastest-growing microblogging and online social networking services. Messages posted on Twitter have been reporting everything from daily life stories to the latest local and global news and events. Monitoring and analysing this rich and continuous usergenerated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. The author in this paper provides a survey of techniques for event detection from Twitter streams. These techniques aim at finding real-world occurrences that unfold over space and time.

**"Using TF-IDF to Determine Word Relevance in Document Queries Juan Ramos"** In this paper, we examine the results of applying Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in a corpus of documents might be more favorable to use in a query. As the term implies, TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in, suggesting that if that word were to appear in a query, the document could be of interest to the user. The author has provided evidence that this simple algorithm efficiently categorizes relevant words that can enhance query retrieval.[5 ]

## 5. Architecture

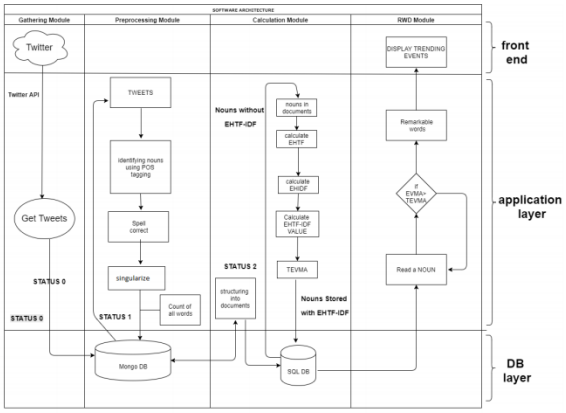


Fig 5.1: System Architecture

- **MYSQL:**

MySQL is the most popular Open Source Relational SQL Database Management System. MySQL is one of the best RDBMS being used for developing various web-based software applications. MySQL is developed, marketed and supported by MySQL AB, which is a Swedish company. MySQL is the most popular Open Source Relational SQL Database Management System. MYSQL Enterprise edition includes the most comprehensive set of advanced features & management tools for MYSQL. MYSQL is the world's most popular open source database. Whether you are a fast-growing web property, technology ISV or large enterprise, MYSQL can cost-effectively help you deliver high performance, scalable database applications MYSQL is popular choice of database for used in web application & is a central component of widely used LAMP open source web application software stack. MYSQL Query Analyzer: To optimize performance by visualizing query activity and fixing problem SQL code.

- **Python:**

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python is Object-Oriented-− Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

Python is a Beginner's Language − Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

- **MongoDB**

MongoDB is an open-source document database and leading NoSQL database. MongoDB is written in C++. This tutorial will give you great understanding on MongoDB concepts needed to create and deploy a highly scalable and performance-oriented database.

## 6. Overall Description

### 6.1 PRODUCT PERSPECTIVE:

With the help of this paper, system can save time and maintain security for fund transfer. The system gives all the information about the e-shopping to provide better service for the customer. It provides facility to the customer to pay by online transaction entering account details, it provides the facility to the customers who want to shop online and online payment due to lack of time. Usage of Steganography ensures that the CA does not know the customer authentication password thus maintaining customer privacy.

### 6.2 REQUIREMENTS:

### SOFTWARE REQUIREMENTS:

Programming language: Java, JSP
Database requirements: MYSQL
Software requirements: Eclipse Kepler, Apache Tomcat Server 7, JDK 1.7

**HARDWARE REQUIREMENTS:**

1) 250 GB HD
2) 2 GB RAM

## 6.3 PRODUCT FUNCTION:

1. EXTRACTING TWEETS For extraction of Tweets we have used Tweepy. It is a twitter API used to fetch the tweets easily into a Python code. Tweepy also has features to manage your twitter account. For getting access to the tweets, first we need the twitters permission. For this, we need to register our application with twitter as a developer, then twitter provides us 4 sets of keys

- Consumer Key
- Consumer Secret Key
- Access Token Key
- Access Token Secret Key This kind of Authentication is called Oauth. Then we make a query upon the text (candidates name) entered by the user and then extract the tweets with the query. The results returned are in JSON format. The result returned contains- Tweets username, retweets count, timestamp. Tweettext, tweet followers count, source, geo-location, mentioned count, etc.

2. PREPROCESSING TWEETS Here we are filtering the tweets fetched by tweepy to extract the relevant data i.e nouns from the tweets. These nouns are then used for event detection. For cleansing data, these steps are performed:

- Translation to English if the tweet is not in English. Noun extraction using POS tagging.
- Spell check if performed.
- Word singularization.
- Lower casing the nouns.
- And finally, storage.

3. STORING TWEETS For storing of the tweets we have used MongoDB as our backend. The tweets in our MongoDBare stored in various stages which are marked by their status. These stages are :

Stage 1:Tweets are fetched from twitter server and stored directly.(status =0)

Stage 2:Status 0 tweets are pre-processed and stored in MongoDB.(status =1)

Stage 3:Status 1 tweets are fetched in form of chunks(doc), their nouns are stored in SQLdb accordingly.(status=2) Features :

- Replication High Availability
- Document-Oriented Storage
- High Performance
- Easy Scalability
- JSON

The main idea behind choosing it is because its a NOSQL database with no restrictions on the size format of the data and also it stores the result in JSON format which is also the format of the results returned by Tweepy. We only save tweets timestamp, tweets text and tweets id.

**7. Calculations:**

The Extended Hybrid TF-IDF, shorted as EHTF-IDF, is an algorithm to quantify importance of words included in a tweet. The algorithm is constructed by two calculations, which are Extended Hybrid TF and Extended Hybrid IDF. They are shorted as EHTF and EHIDF respectively. An EHTF value is gotten based on appearance frequency of a word among current tweets. If the value is high, the word can be considered as an important word. On the other hand, an EHIDF value is gotten based on generality of a word among tweets. The tweets include not only current tweets but also past tweets. In this calculation, if the tweets include many numbers of one word, an EHIDF value of the word is low. It means that a generally appearing word among the tweets is not important, since a word of this kind is a generally used word. According to EQ.1, EHTF-IDF quanties importance of a word multiplying an EHTF value of the word by an EHIDF value of the word. This multiplying is calculation to give high importance to a word which frequently appears in a current tweets but is not a general used word. EHTF-IDF = (EHTF)∗ (EHIDF)...(EQ.1)

1) Extended Hybrid TF: First all current tweets are organized into a document. Second count a total number of appearance of all words in the document. Third count a number of appearance of a word in the document. Finally an EHTF value of the word is gotten by dividing the number of appearance by the total number. This calculation is as the following equation. EHTF = (NAW) ÷(T NAAW)...(EQ.2) Where, NAW is a number of appearance of a word in a do5cument, and TNAAW is a total number of appearance of all words in a document.
2) Extended Hybrid IDF: In the proposed algorithm, all current tweets are organized into a document as with the EHTF calculation, and then each past tweet is respectively considered as past documents. The past documents are used to evaluate generality of a word which included in the current document according to EQ.

3) On the other hand, in Hybrid TF-IDFs, each current tweet is respectively considered as current documents and past tweets are not used. The Hybrid IDF calculation uses these documents to evaluate generality of a word which included in them. In this way, it is difcult that this processing works effectively. It is the reason why the proposed method is more accurate than Hybrid TF-IDF. EHIDF = log [(T NP D) ÷ (NP D)] + 1...(EQ.3) Where TNPD is total number of documents and NPD is number of documents including that word which is included in current document.

4) RWD MODULE: Degrees of importance of words can be represented in numerical form by EHTF-IDF. As it is no more than numerical values, it is impossible to detect remarkable words only from the numerical values. Therefore threshold is required for the detection. The Remarkable Word Detecting Method, shorted as RWDM, is designed for dynamically getting the threshold and then detecting remarkable words efciently. This method uses two moving averages, which are EHTF-IDF Value Moving Average and Total EHTF-IDF Values Moving Average, respectively shorted as EVMA and TEVMA. The reason why the proposed method uses the moving averages is to reduce a rate of false detection.Sometime,an EHTF-IDF value includes noise. As a moving average is less affected by noise, this method uses EVMA and TEVMA. In this method, TEVMA is used as threshold, so a word can be judged important if EVMA of the word exceeds TEVMA.

5). DISPLAYING RESULTS :Once the remarkable words have been detected from tweets, they are displayed as events.

## 8. Advantages:

- Twitter is fast .
- When you publish on it.
- The people following your business will receive the tweet immediately.
- The business owners can make use of this feature to test the different campaign angles that they are considering.

## 9. Conclusion

Development of this project titled Twitter Based Event Detection has been a useful experience. We gained knowledge about the APIs, Python, TextBlob and various other technologies and platforms. Weve learnt new concepts of working. The project aims to demonstrate how Twitter Data Analysis can be used to give a clear picture about a particular event, so that the user can get an overall idea about events happening currently. The project was analyzed, designed, developed, tested and deployed successfully.

## 10. References

[1] Koichi Sato, Junbo Wanga and Zixue Cheng, ”Detecting Real-time Events using Tweet” 2016 IEEE Symposium Series on Computational Intelligence.

[2] Rui LI, Kin Hou Lei, Ravi Khadiwala and Kevin Chen-Chuan Chang, ”TEDAS: a Twitter Based Event Detection and Analysis System”, 2012 IEEE 28th International Conference on Data Engineering.

[3] Farzindar Atefeh, Wael Khreich,"A Survey of Techniques for Event Detection in Twitter ", published by - Wiley Periodicals, Inc,September 2013 .

[4] Kumar Shamanth, Liu Huan, Mehta Sameep, L. Venkata Subramaniyam, " Exploring a Scalable Solution to Identifying Events in Noisy Twitter Streams", Proceedings of the 2015 IEEE/ACM InternationalConference on Advances in Social Networks Analysis and Mining, 2015.

[5]"Using TF-IDF to Determine Word Relevance in Document Queries ",by Juan Ramos, January 2003.

[6] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors", Proc. 19th InternationalConference on WWW2010, 2010.

[7]1. D. Inouye and J. K. Kalita, "Comparing twitter summarization algorithms for multiple post summaries", 2011 IEEE InternationalConference on Privacy Security Risk and Trust and IEEE InternationalConference on Social Computing, 2011