To Study, Analyze and predict the Diseases using Big Data

Archana Bakare
ME (CSE) Student, Department of Computer Science & Engineering, Walchand Institute of Technology, Solapur, Maharashtra, India

Rajesh Argiddi
Associate Professor, Dept. of Computer Science and Engineering, WIT, Solapur, Maharashtra, India

ABSTRACT

As we know in today's life Twitter, Facebook, Google plus are well known social media now that user can use this application for different purposes. Nowadays many people have many social accounts Twitter is an online news and social networking service where users post and interact with messages, "tweets," limited to 140 characters. Registered users can write tweets, but those who are unregistered can only read them. Users access Twitter through its website or a mobile device app. Twitter is one of the growing social site that people are using for connecting, sharing with each other. There is number of short text messages posted by many people called as tweets. It is very hard to do analysis of social media data which has large amount of noisy, informal text of tweet messages and stream data.

General Terms

K-means algorithm used for Classification,Multi-rank Algorithm for clustering.

Keywords

Data mining, social media, prediction.

1. INTRODUCTION

1.1. Motivation

Data mining has several techniques to find and analyze data from different sources and shortly it into useful information. Usually the Internet provides large of information than required. Even though cleaning process is applied on text, processing and summarizing this text, small messages for human beings is difficult. A Social media offers best platform to people for communication and interacting, status update, way to express thoughts regarding many topics. Twitter is one of the familiar social site on which people continuously search for trending topic, especially when surfing the internet with their mobile devices which have too small screen than computer screen. Twitter that allow their users to update, comment, and connect with users in their social circle.

Twitter users can write and read short length 140-character messages called as tweets. Many People usually upload their everyday routines, update status, other events on social sites.

1.2.Problems and solution

Whenever extracting information from social media such as the large amount of fast arrived stream of data from Twitter, researchers and analysts face some issues such as language difficulty, short and unstructured message, some tweet contains only hyperlinks and so on. In the case of social sites, daily 500 million short tweets are posted on twitter in a different languages. It is difficult to extract key contents of topics from social sites and becomes time consuming task to read millions of tweets and to find the reasonable tweets. For this we only extract tweets related Asthma ,Blood pressure & Diabetes then apply streaming. Whenever get tweets from twitter for easily read that tweets, we apply minimization stream, for reasonable tweets, then apply K-means algorithm to classify the diseases.

2. RELATED WORK

In this presented algorithms for k-means after applying minimization of tweets of micro blog posts. Algorithm is used for collections of short posts or messages on specific topics that are available on the social site Twitter and displayed with groups of messages on that specific topic. Here, goal was to produce prediction of result & their accuracy of collection of posts, messages on that particular topic.[1]

There has been increasing interest in gathering non-traditional, digital information to perform disease surveillance. These include varied datasets such as those stemming from social media, internet search. Twitter is an online social media platform that enables users to post and read 140-character messages called “tweets”. More importantly, tweets are often tagged by geographic location and time stamps potentially providing information for disease surveillance [2,3]

Social media have been predicted as a data source for flu surveillance because they have the potential to offer real-time access to millions of short, physically localized messages covering information regarding personal wellbeing. However, accuracy of social media surveillance systems declines with media attention because media attention increases “chatter” – messages that are about flu but that do not pertain to an actual infection – covering signs of true flu prevalence. This paper précises that recently developed flu infection detection algorithm that automatically categorizes relevant tweets from other chatter, and describe current flu surveillance system.[4]

In this System is used and presented the real-time interaction of events in Twitter. K-Means algorithm for monitoring tweets to detect particular diseases. For detecting particular diseases, System made use of classifier of tweets streams which is based on special features such as keywords presents in a tweet, the total number of words present in tweet and their context. They also considered as every user of twitter can be and apply filtering and particle filtering. Both are used for detection of location .

This proposed method based on analysis of messages posted on the micro-blogging site Twitter.com to determine if a similar connection can be discovered. They proposed several methods to identify influenza-related messages and equate a number of regression models to correlate these posts . Using over 500,000 messages spanning 10 weeks, we find that our best model achieves a correlation of .78 by leveraging a document classifier to identify relevant messages. Analyzing user messages in social media can measure different population features, including public health measures. For example, recent work has correlated Twitter messages with influenza rates in the United States; but this has largely been the extent of mining Twitter for public health. It considered a broader range of public health applications for Twitter. They introduced postponements to include prior knowledge into this model and apply it to several tasks: tracking illnesses over times (surveillance), measuring behavioral risk factors, localizing illnesses and analyzing similar and medication usage. Their results suggest that Twitter has broad applicability for public health research.[5]

In this paper provides a review of different approaches for extracting information from the Social Web for health information: a health approach. Also different approaches for extracting information from social web applications to personalize health care information. The model we use in this paper could be used to analyze tweets for health care personalization. Finally, the public is considering the larger impact of how social media can impact health care, where patients can “friend” doctors and always share information among thousands of friends [6]

This paper presented algorithms for k-means after applying minimization of tweets of micro blog posts. Algorithm is used for collections of short posts or messages on specific topics that are available on the social site Twitter and displayed with collections of messages on that specific topic. Here, goal was to produce prediction of result & their accuracy of collection of posts, messages on that particular topic.

3. METHODOLOGY

The following figure describes designed framework of the system. The goal of designed system is to generate summaries of different topics. The overall architecture of system is shown in figure 1.The system first fetches tweets from Twitter using Twitter API stream library.

3.1. Tweet collection module:

This module helps to get tweet stream from twitter by using twitter stream API. In this module, we used Twitter4j library for getting continuous stream of tweets. Tweet filter is used to remove links, non-English, symbols,retweet sentences. The filtered data is stored in a file system.

3.2. Classification module

K-means clustering is a **data mining**/machine learning **algorithm** used to cluster observations into groups of related observations without any prior knowledge..

This can be used to classify tweets into disease categories (Asthma, Blood pressure, Diabetis). Training data is number of tweet files from the data set used by Multirank walk classifier for training. Training data is provided in tweet files and classified in to the general categories

3.3. Algorithm:

First we take tweets from twitter after that cleaning process we applies k-means algorithm.

When applies k-means clustering algorithm diseases wise categorized ,like asthma related tweets are on one cluster, Blood pressure related tweets &Diabetes related tweets.

On this we applies Multi rank walk algorithm suppose we applies on Diabetes then it will classifies as Type 1 & Type 2 Diabetes ,Then we applies on Blood Pressure then it will classify as Low & High Blood Pressure.

First initialize the classifier by using category array.

Training data is stored and organized into the category and it is read from the twitter database. After reading data, resulting data is used to train classifier for each and every Category. Next for each category, read all testing data and execute Multirank walk algorithm

This process is continued till the end of all the categories.

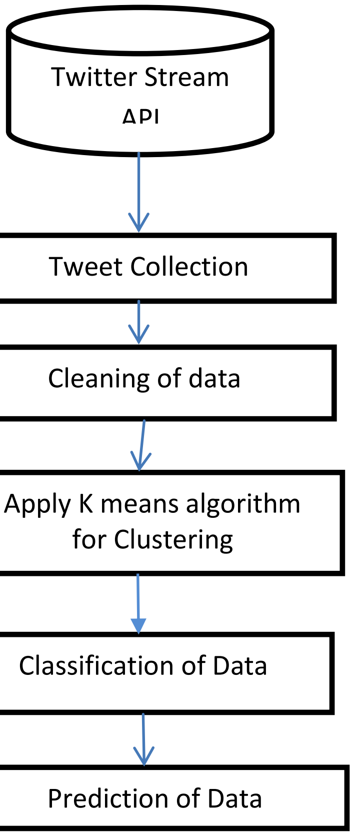

Figure.1.System Architecture

3.4. Result:

A result of this paper which will predict the fore-coming strokes of different diseases depending on data gathered by social networking sites. The impact of such diseases can be minimized by predicting the future stroke timing. It can be possible by studying the different types of data collected over those diseases. The results can be helpful for public health investigation and targeted patient interventions.

4. EXPERIMENTAL SETUP AND RESULTS

For making use of twitter data, the developer account is needed. With the help of twitter4j library system collects the tweet stream from Twitter site. Then stream of tweets is given to K-means Algorithm, the algorithm uses training data to classify tweets in to the specific categories for topic detection. Results of classification are in the form of general categories.

For classification results, experimentation starts with collecting tweets on topics such as Asthma, Blood Pressure, Diabetes using Twitter Streaming API.

Table 1 presents results sets which show the overall classifier accuracy for the classifier performance and multirank walk algorithm can reach more than 60% of classification correctly.

Table 1: Classification results

Sr.no	Category	Total tweet count	Tweets used	Accuracy %
1.	Asthma	300	195	65
2.	Diabetes	300	200	66
3.	Blood pressure	300	190	63

5. CONCLUSION AND FUTURE WORK

In this paper, implemented k-means algorithm & Multirank-walk algorithm used, on tweet streams to provide useful twitts,apply filtering technique with regard to topics. Then whenever user searches information on particular diseases that can help the user to get an overview of diseases quickly. The implemented work determines,

Categories of the tweets and classifies them into different Categories using Multi-rank walk algorithm that provides Accuracy approximately 60% to 70% and gives the result to the user. System removes short, dis-similar and noisy Nature of the tweets using classification. The future work is this will be on mobile.

6. REFERENCES

[1] Predicting Asthma-Related Emergency Department Visits Using Big Data Sudha Ram, *Member, IEEE*, Wenli Zhang, Max Williams, and Yolande Pengetnze, *MD*

[2] Broniatowski, David A., Michael J. Paul, and Mark Dredze. "National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic." PloS one vol. 8, no.12, e83672, 2013.

[3] Kim, Eui-Ki, et al. "Use of hangeul twitter to track and predict human influenza infection." PloS one vol. 8, no.7, e69305, 2013.

[4] The Twitter of Babel: Mapping World Languages through Micro blogging Platforms Delia Mocanu, Published: April 18, 2013

[5] Culotta, Aron. "Towards detecting influenza epidemics by analyzing Twitter messages." In Proceedings of the firstworkshop on social media analytics, pp. 115-122. ACM, 2010.

[6] By L Fernandez-Luque - 2011 Review of extracting information from the Social Web for health personalization. Fernandez-Luque L(1), Karlsen R, Bonander J.