Analysis of Health-Tweets using K-means clustering

Shalini L[1], Gopali Naga Sravya[2]

[1]*Assistant Professor (Senior), Dept. of Computer Science and Engineering, VIT University, Tamilnadu, India*
[2]*Student, Dept. of Computer Science and Engineering, VIT University, Tamilnadu, India*

Abstract - *Analysis of tweets plays an important role in understanding the present situation in society and the opinion of people towards the situation. These tweets are useful for addressing the challenges faced by the current situations in the society. These tweets also provide a lot of information about the thinking and behavior of the individual in the society. The tweets can be related to any present topic like politics, health, business, marketing, and online trends etc. Analyzing these tweets can be useful for understanding the present situation related to this topics. In this paper, we analyzed tweets from several records and using several sources like bbc health, cbc health, cnn health, everyday health, fox news health, gdn health care, good health, Kaiser health news, latimes health, msn health news, nbc heath, npr health, nytimes health, Reuters health, us news health, wsj health. So, by combining the tweets from these 16 records, top 5 clusters of positive and negative words are obtained by using K-means clustering, the analysis is done on the number of positive and negative words present in those records, and the corresponding plot for the frequency of top 40 words is obtained. The purpose of this analysis is to find the views of different people related to the news on health in different channels.*

Key Words: tweet analysis, health tweets, K-means, clustering, positive sentiment, negative sentiment, frequency of words.

1. INTRODUCTION

The data can be obtained from several sources of social media like Twitter, Facebook, Instagram etc from individuals, neighborhoods, and groups. The tweets can be done on the images, links, any questions or quotes posted by any other person, and re-tweets can also be done on the other tweets posted by some other person. Tweets can also be composed of daily conversations, movies, news, life, politics, and updates etc. We can also use the tweets related to the medical field and analyze the tweets, for finding how the health will change over time based on several other patterns. The approach here consists of preprocessing of data, stemming of the data, removing stop words, classifying the words into positive, neutral, and negative sentiments, getting unique words for each sentiment, finding the frequency of each word under respective sentiment, plotting the frequency of the words, and clustering of the words by using K-means algorithm. In this analysis, positive tweets are presently more than the negative tweets, which mean that the positive sentiment is more related to the health tweets collected, compared to the negative sentiment.

The day, month, year, and time of the tweet is also present in the records which can be useful for analyzing the tweets related to the health based on the time, and how it improved or not improved according to the time. So, by knowing the standards of the treatment given to a particular disease according to the time, we can also improve the treatment given to the people based on their opinion. This analysis can be used to solve the problem of the people related to their health issues and it can also be useful for the improvement of the treatment given to them based on positive and negative tweets. This process of extraction of information which is either known or known from a large number of datasets is known as Data Mining. Retrieval of information, analysis of quality in the treatment can also be done using this process.

2. LITERATURE SURVEY:

Surveillance of health is the important task of tracking the happenings that are related to the health of human beings and one of the important areas is the pharmacovigilance. It tracks and monitors the safe usage of pharmaceuticals products. This involves the tracking of side effects which are caused by drugs related to health and medicines. Collecting this information by the medical professionals has a difficult time. It contains a lot of noise in the data like special symbols, hyperlinks, stop words, punctuations etc. Removing this unnecessary information and applying Deep learning algorithms including the neural networks for classifying the information, that can be helpful for detecting the personal experience in the form of tweets is done in this paper and finally, they produced a method called Deep Gramulator which improves the results [1].

The social media can provide a large amount of data for this purpose. Twitter data can be used for the purpose as it provides us a majority of the information when compared to other networks. They analyzed about 12,000 tweets given by nearly 10 classes of users by the year 2010. The purpose of analysis includes the healthcare personalization of individual, health disparities are discovered, advertising and monitoring of public health. The scores are calculated in order to indeed a good marker of the health of the user [2].

A survey was done on big data analytics in the healthcare domain and the methods used, technology and algorithms used for the big data are analyzed, for the purpose of decision making and management in taking care of health. In this survey paper, the outcome will prove that this survey is beneficial to the academician, industries, and researchers who have interest in the big data analytics and health care [3].

A three-stage filtering, which also includes the process of categorization and with three different types of knowledge resources using the NLP (Natural language processing) for filtering and also extracting personally relevant and articles of health-related news which are referred in the tweets of Twitter. This filtering is done because of a large number of posts present in the social media networks and also a large number of articles which has the news related to the health. So, using the three stages of term based filtering, filtering of content and categorization large volume can be analyzed [4].

Diffusion of tweets can occur because of tweeting and re-tweeting did by the users of the twitter because of follower friend connections. Even if the users won't follow the other users also they can re-tweet to their tweet. A study was conducted on results to quantify the diffusion of health-related tweets through the re-tweeting activity. This includes the development of a graph framework related to the re-tweets and an analysis this graphs and diffusion associated metrics for tweets related to health. This is the first attempt done on the study of health information analysis based on the re-tweet graph analysis [5].

Several tweets will occur in Twitter based on the certain topic during the occurrence of tweets. But among these tweets, not all the tweets are related to the topic and are important. Some of the tweets are not of importance to that topic. So, classification of tweets based on the interest of the individual is important in some cases. The objective of classifying the tweets into two categories of "like" and "dislike" is based on the selection of features and also from considering the preferred news for the topic. The outcome of the experiment will show that a training done on the less number of tweets will give the personalization for the next incoming tweets [6].

Depression related analysis of tweets was done for finding of the mental health and thinking of the people. They assessed themes by using a random sample of n=2000, tweets related to depression. These tweets were coded related to the expression of DSM-5 symptoms and for MDD (Major Depressive Disorder). Their findings can be useful for the health professionals, to tailor, prevention of the target, and sending messages of awareness to the users of twitter whom they find as depressed [7].

Collecting several opinions from the people from the different famous websites in which patients will express the opinions, experiences they faced, and side effects caused by the drugs used on them for the treatment. Classification of users is done based on the opinion given by them on the drug. Knowledge-based discovery is also done on the people and patterns are found based on symptoms, drug, and disease by using Association rule mining [8].

Overwhelming of tweets will occur if the user follows many accounts when they interested in the content's subset. In order to overcome this problem of overwhelming, filtering of the incoming tweets is done which is based on the interest of the user. A classifier is proposed for this purpose and the evaluation of tweets by this classifier will get an accuracy of 89% [9].

Over 100,000 tweets related to the US presidential election which is taken place in the year 2012 and several attributes are also present, are collected for the analysis of sentiment of tweets, emotion of the people in the tweets, purpose of the different features in the specific topic, and style in the tweets. Development of several classifiers for the different analysis of these tweets like described is done. The results obtained will provide a baseline for the automatic systems which can be used further on the new data [10].

3. METHODOLOGY:

Acquisition analytics can be done by using the clustering or segmentation techniques. One of the most important techniques used for clustering of data is K means clustering. It is used for analysis of unlabelled data, which is called as unsupervised learning. Structure of the underlying data can be found using this unsupervised learning, instead of trying to optimize for a specific and labeled data. The relationship between the words is formed and then the analysis of the data can be done using the clustering technique.

In the K means clustering the data can be clustered into groups based on the similarity in the observations of each cluster that can be able for extracting the insights from a large amount of unstructured data which can be collected from several sources. Whenever we want to analyze the large amount of data obtained in form of comments or tweets, from Twitter, Facebook or YouTube, for a particular topic, it is not possible to analyze the views of different people by looking manually into the comments given by them. So, here sentiment analysis of the tweets comes into the picture.

The fundamental thought of K means clustering is first forming a K seeds (which is one of the methods of data partition) and next grouping is done based on the observations of the K clusters which can be calculated by using the distance measurements with each of the K seeds. Text data that is used for analysis can be converted to a form of numeric data for the clustering algorithm to be applied. The grouping of the observation with the already formed clusters will be done if the distance between the observation and the seeds in the already formed cluster is the minimum one when compared to the distance between the observation and other seeds in the other formed clusters.

The steps are done on the data and the results obtained are discussed in the following sessions.

4. IMPLEMENTATION STEPS AND RESULTS

(i) First the 14 documents (like bbc health, cbc health, cnn health, everyday health, fox news health, gdn health care, good health, Kaiser health news, latimes health, msn health news, nbc heath, npr health, nytimes health, Reuters health, us news health, wsj health), which have different tweets, which also includes the time, date, and the hyperlink of the website where the tweets are done on the news related to the tweets.

(ii) These tweets are combined into a single document.

(iii) Cleaning of the data can be done by removing the numbers, hyperlinks, special symbols, single and double character words, which will be of no use for the analysis.

(iv) We have to import the stop words and word tokenize from the NLTK (Natural language process toolkit) corpus, for the tokenization of words, removing the stop words, stemming of the words, and filtering of the words. Porter stemmer is used for the purpose of the stemming the words. It is easy to get the positive and negative words clusters from the data, by doing this process.

(v) The sentiment of each word should be found by using the sentiment analyzer which can be imported from the same NLTK corpus. Based on the polarity of the word, the corpus of positive words, negative words, neutral words can be formed.

(vi) The bigrams of the different sentiment like positive, negative, neutral are also obtained from the data.

(vii) Term document matrix should be obtained for the words for further process of clustering using K means. Term frequency can be found out which can be used for clustering purpose. Here, 5 clusters are made for the words, by fitting the model with the term frequency vectorizer of the words. The clusters can be made for the positive words and negative words separately so that we can avoid the diffusion of both clusters. The 5 clusters for positive and negative words are shown below:

- Positive words cluster:

Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
yummi hurrah glee generos gener gain funniest freedom free fascin	award yummi fascin glee generos gener gain funniest freedom free	happi yummi glori generos gener gain funniest freedom free fascin	luckili yummi glori generos gener gain funniest freedom free fascin	glori energ glee generos gener gain funniest freedom free yummi

Table-1: Clusters of positive words

- Negative words cluster:

Cluster0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
worst enemi fiasco feloni fear fatal failur fail excruci evil	poor worst disappoint fear fatal failur fail excruci evil enemi	bank worst dumb feloni fear fatal failur fail excruci evil	contempt worst fiasco fear fatal failur fail excruci evil enemi	alarm worst drown feloni fear fatal failur fail excruci evil

Table-2: Clusters of negative words

(viii) Frequency plot of the positive and negative words are obtained by considering samples on X-axis and their counts on Y-axis. This frequency of words is used for finding the topic on which the tweets are related to. So, this is used for finding the information from the data of unknown topic.

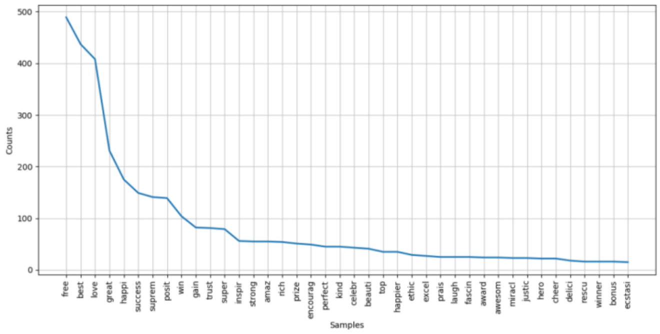

Fig -1: Frequency plot of positive words

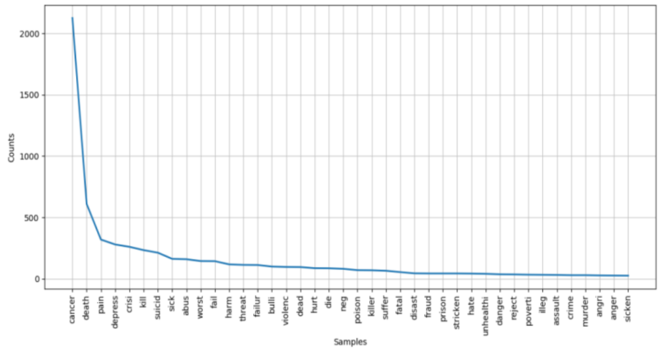

Fig-2: Frequency plot of negative words.

Top 40 words frequency plot is represented in the above figures. By using the frequencies of the words in the data it will be useful for finding the topic on which the tweets are made by the users of several different websites.

5. CONCLUSION:

By using clustering, it will be useful for finding the group of words with similarity. It is also useful for finding the sentiment of the words without any labels that are it is related to unsupervised learning. Words are divided based on their sentiment label, which can be found by using the NLTK corpus. K means clustering is done on the words which are divided that are present in the document, which is formed by combining the 14 records. The frequency of words is also found which is useful for finding the topic on which the tweets are made by the users of several different websites. This will be useful for getting the information about the topic. The new tweets are clustered based on the similarity between the newly occurring words and with the words that are already present in the clusters which are done by applying K means clustering algorithm on the preprocessed words.

REFERENCES

[1] Calix, R. A., Gupta, R., Gupta, M., & Jiang, K. (2017, November). Deep gramulator: Improving precision in the classification of personal health-experience tweets with deep learning. In Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on (pp. 1154-1159). IEEE.

[2] Kashyap, R., & Nahapetian, A. (2014, November). Tweet analysis for user health monitoring. In Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on (pp. 348-351). IEEE.

[3] Kumar, H., & Singh, N. (2017). Review paper on Big Data in healthcare informatics. International Research Journal of Engineering and Technology.

[4] Steele, R., & Min, K. (2012, July). Health news feed: Identifying personally relevant health-related URLs in tweets. In Industrial Electronics and Applications (ICIEA), 2012 7th IEEE Conference on (pp. 491-496). IEEE.

[5] Bakal, G., & Kavuluru, R. (2017, February). On quantifying diffusion of health information on Twitter. In Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on (pp. 485-488). IEEE.

[6] Weilin, L., & Hoon, G. K. (2015). Personalization of trending tweets using like-dislike category Model. Procedia Computer Science, 60, 236-245.

[7] Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., & Bierut, L. J. (2016). A content analysis of depression-related tweets. Computers in human behavior, 54, 351-357.

[8] Prashanth, M. V. (2016). Opinion Mining and Summarization of User statements in Health Communities.

[9] Altammami, S. H., & Rana, O. F. (2016, November). Topic Identification System to Filter Twitter Feeds. In Soft Computing & Machine Intelligence (ISCMI), 2016 3rd International Conference on (pp. 206-213). IEEE.

[10] Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. Information Processing & Management, 51(4), 480-499.