International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

Mining of Health-Related Topics using Twitter

V. Karpagam[1], K. R. Aiswarya[2], R. Chella Barghave[3]
[1]*Assistant Professor, Department of CSE, K. L. N. College of Engineering, Madurai, India*
[2,3]*UG Student, Department of CSE, K. L. N. College of Engineering, Madurai, India*

***Abstract*: Social media has become a major source for analyzing all aspects of daily life. It holds huge amount of data and our goal is to obtain value from the massive amount of data present in social media. Twitter is one of the prominently used social media platform. In this paper, we are going to use the concept of Data Mining to monitor people's health using the tweets tweeted by the people in twitter. Data Mining helps in finding the hidden patterns and correlations within large data sets to predict the outcome. Public health can now be observed on 'Twitter'. These tweets are collected based on two parameters namely, 'Time' and 'Geographical Location'. Time indicates the tweets that are collected in recent years and geographical location indicates the tweets that are collected from all over the world. We use K-means clustering algorithm to cluster the data and identify the rate of influence of various diseases which has major impact on people over time. In this algorithm, K indicates a positive integer number. This partitioning clustering technique is the most popular and effective technique. K-means clustering aims to partition 'n' observations into k-clusters in which each observation belongs to the cluster with nearest mean, serving as a prototype of the cluster. Generally, we fix the number of K-clusters and the grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. It is relatively faster clustering technique which has wide range of applications. K-means clustering algorithm is an unsupervised learning algorithm that acts well on unlabeled data-sets and provides results with more accuracy.**

***Keywords*: Clustering, Data Mining, K-means algorithm, Twitter.**

1. Introduction

The concept of 'Data Mining' plays a vital role in today's technological world. Data Mining can be defined as a process used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. It also includes extraction of interesting patterns, previously unknown and potentially useful information from huge amount of data. There is a famous saying that, "We are drowning in data, but starving for knowledge!". The solution to this problem is brought to light by the concept of data mining.

Social media has become part and parcel of our everyday life. Most of the people share their personal and general information in social media such as facebook, twitter, instagram etc. If people go somewhere or go to a new destination, they upload their status in social media from time to time. People also express their emotions through social media. If they are happy, sad or sick, they upload their status in social media to make sure their friends, family and relatives are aware about them. Storing huge amount of data in a database is a challenging task and we have big data technology for that purpose. Big data are larger and complex data sets. They are so voluminous and we cannot manage them using traditional data processing software. Actually these massive amount of data can be used to address business problems that we wouldn't have been able to tackle before. The three V's related with big data are Volume, Velocity and Variety. Big data technology makes it possible for us to gain more complex answers because we have massive information. In general, complete answers mean more confidence in the data, which is a different approach in tackling problems.

Social media holds huge amount of data and it would be meaningful, if we are able to obtain value from those data. The main objective is to turn data into information and information into insight. We obtain data of various kinds from social media but here we take specific datasets to analyze the hidden information and bring out the true values from it. We deal with mining of health-related topics from twitter. Twitter holds various data such as text, image, audio, video, Graphics Interchange Format(GIF) and all other kinds of multimedia too. We mainly concentrate on mining of text from twitter using the hashtags used by the people. Social media data mining can provide even more powerful intelligence from the information gathered from social media. Social media mining uses a range of basic concepts from computer science, data mining, machine learning and statistics.

2. Objectives of the project

The main objective of the project is to mine health-related topics using one of the prominent social media "Twitter" with the help of the tweets tweeted by the public. It is done based on two factors and they are "Time" and "Geographical Location". By doing so, we tend to obtain the numerical percentage of various diseases and we can filter out the top trending diseases in recent. We would be able to make people aware about those diseases which have high rate of influence on people and the awareness can be done by conducting seminars and workshops and lot more.

3. Existing system

Social media has become a major source of information for analyzing all aspects of daily life. Ailment Topic Aspect Model (ATAM) was used to observe public health on twitter. The use of tweets literally has several benefits including instantaneous data availability at virtually no cost. Early monitoring of health data is complementary to post-factum studies and enables a range of applications such as measuring behavioral risk factors and triggering health campaigns.

We formulate two problems: health transition detection and health transition prediction ATAM algorithm is an extension of Latent Dirichlet Algorithm (LDA). Latent Dirichlet Algorithm explains non-ailment topic but also includes a model to filter out background noise and a specialized ailment model that incorporates symptoms and treatment information. For example, consider that a user tweeted in his/her twitter page as, “Damn flu, home with a fever watching TV”. This little tweets actually holds a bundle of information. This tweet is categorized in various categories according to the algorithm. Here the words related to ailment are {flu, fever} and the other words that are topically related are {home, watching, TV} which might be described by “stay at home” topic. There are also common words that would not be described with a particular topic or ailment such as {damn, with, a}.

Ailment Topic Aspect Model (ATAM) designed to uncover latent health-related in a collection of tweets. The proposed method achieves remarkable improvement over LDA. Its novelty is that it distinguishes background words such as “home” and “watching TV” from health related words such as “hurts” and “allergy”. For each document, these health related words are considered to correspond to a unique ailment such as “obesity”, “insomnia” or “injuries”.

- Here each tweet d is categorized with an ailment ad= i with probability of ni.
- Each word token n in tweet d is associated with two observed variables.
- The word type Wdn
- A label Ydn that we call the aspect denotes whether the word is a symptom word, treatment word or anything else such as general word.
- The Y variables are given as input.

The dataset is labelled using the list of 20,000 symptoms and treatment key phrases. Each word token in a tweet is generated based on background model, topic model and ailment model.

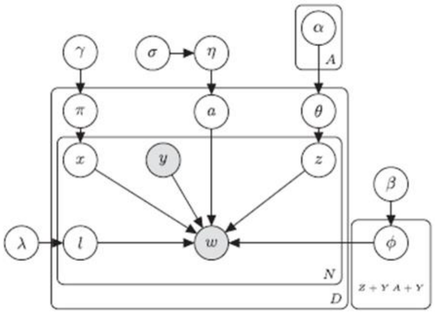


Fig. 1. Ailment Topic Aspect Model

ATAM algorithm is further divided into TM-ATAM and T-ATAM. Here TM-ATAM stands for Temporal Ailment Topic Aspect Model and T-ATAM stands for Time Aware Ailment Topic Aspect Model.

- TM–ATAM, a model able to detect health-related tweets and their evolution over time and space. TM–ATAM learns, for a given region, transition parameters by minimizing the prediction error on ailment distributions of pre-determined time periods.
- T–ATAM, a new model able to predict health-related tweets by treating time as a variable whose values are drawn from a corpus-specific multinomial distribution.
- Extensive experiments that show the superiority of T–ATAM for predicting health transitions, when compared against TM–LDA and TM–ATAM, and its effectiveness against a ground truth.

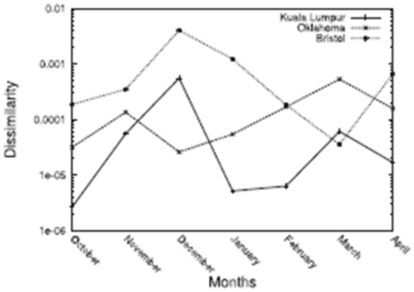


Fig. 2. Topic transitions over time

Time plays a vital role in every aspect of categorization. We collect tweets from all over the world but it’s pertained to certain time limit. The rate of influence of diseases change accordingly to time.

Example:

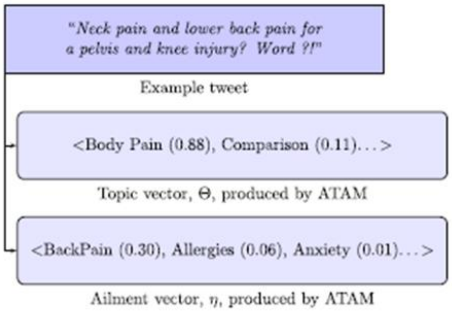

“Neck pain and lower back pain for
a pelvis and knee injury? Word ?!”
Example tweet
<Body Pain (0.88), Comparison (0.11)...>
Topic vector, Θ, produced by ATAM
<BackPain (0.30), Allergies (0.06), Anxiety (0.01)...>
Ailment vector, η, produced by ATAM

Fig. 3. Ө is predicted by ATAM on an example tweet, h predicted by ATAM on all posts containing the example tweet

4. Proposed system

Social Media can be used obtain values from it and our project evolves with the same idea and with the help of famous social media “Twitter”, we are going to mine users tweet and analyze them. Twitter is a platform where people and celebrities share their opinion mostly in words rather than images. Most of the tweets will be in text format and limitations is 160 characters as of now. There will be so many user’s tweet. We are going to monitor people's health over time. Generally, people tweet about health-related topics in certain situation and those instances are as follows.

- When people are actually suffering from the disease then they may tweet about it.
- When people want to know the prevention measure of any disease, then they may tweet about it.
- When their friends or known people suffer from the disease then they may have the idea of tweeting about it. To create awareness about any disease, people may tweet about it.
- When that specific disease is prevalent in their city/state/country/region, then people may tweet about it.

We classify user's tweet based on two parameters and that is "Time" and "Geographical location". To perform all these, we rely on one of the most important concept called "Data Mining". There are so many data mining algorithms to perform clustering and we always need to choose the one that suits our project in all dimensions. As we deal with social media, the amount of data is extremely huge. In our proposed system, we make use of one of the finest and efficient clustering algorithm in data mining called as "K-means Clustering Algorithm".

To overcome the disadvantages of our existing system, we have brought out an efficient clustering algorithm called K-means clustering algorithm which serves the purpose of our project. The major problems related with our existing system are,

- Training large data-sets becomes very difficult.
- Lots of packages exists but scalability becomes the main issue.
- Validation and evaluation is difficult.
- Visualization for the Bayesian model is hard and presenting large data-sets is tedious task.

Thus K-means clustering algorithm helps us in overcoming all these difficulties. This partitioning clustering technique is most popular and effective technique. K-means clustering aims to partition 'n' observations into k-clusters in which each observation belongs to the cluster with nearest mean, serving as the prototype of the cluster. The various steps involved in K-means clustering algorithm are,

- Step1: Input the number of clusters k.
- *Step2:* Initialize the cluster centroid.
- *Step3:* Calculate Euclidean distance.
- *Step4:* Move on to next observation and calculate Euclidean distance.
- *Step5:* Calculate Euclidean distance for the next observation, assign next observation based on minimum Euclidean distance and update the cluster centroids.
- *Step6:* Continue steps 1 – 5 until all observations are assigned.

- Steps involved in clustering the diseases are as follows

Representation:

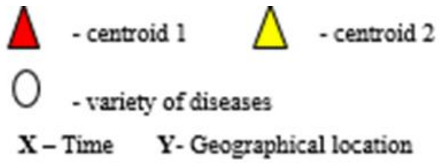


Fig. 4. Representation

Here we consider two centroids to cluster our diseases and as we classify based on two parameters namely time and geographical location we take "Time" in X-axis and "Geographical location" in Y-axis and then we proceed.

Step1:

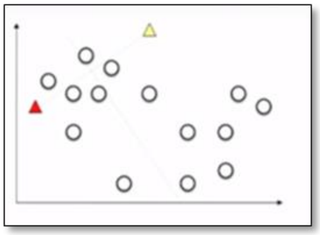


Fig. 5. Dataset holding diseases

Step2:

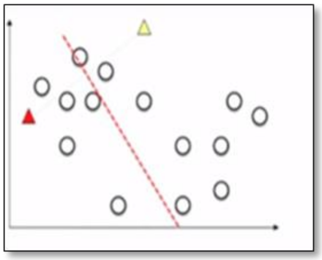


Fig. 6. Centroids defined

Step3:

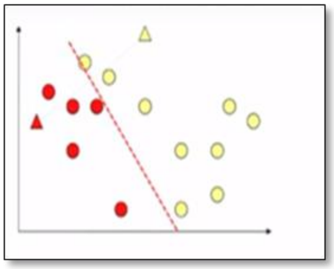


Fig. 7. Iteration 1- clustering

Step4:

We have to move on to next observation and calculate euclidean distance and in the next step we need to calculate euclidean distance for the next observation, assign next observation based on minimum euclidean distance and update the cluster centroids.

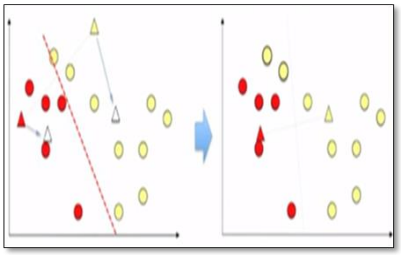


Fig. 8. Centroids gets updated

Step 5:

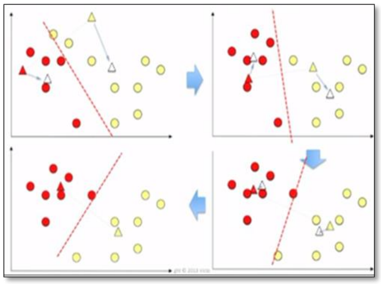


Fig. 9. Various iterations of clustering until all observations are being properly clustered.

Finally, all the observations are clustered in groups and the advantages of proposed system are,

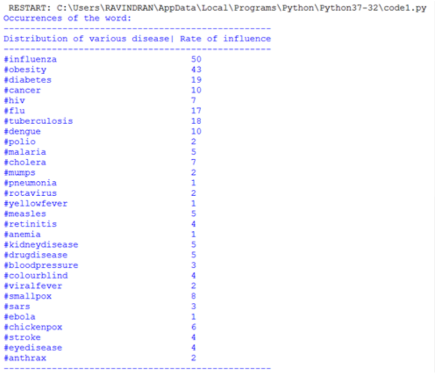
- K-means clustering algorithm is relatively faster clustering technique.
- It works fast with the large data set since the time complexity is O(nkl) where n is the number of patterns, k is the number of clusters and l is the number of the iterations.
- It relies on Euclidian distance which makes it works well with numeric values with interesting geometrical and statistic meaning.
- K-means clustering algorithm is an unsupervised learning algorithm that acts well on unlabeled data-sets and provides results with more accuracy.

5. Applications

In general, when we want to make groups of similar things from a randomly distributed collections, K-means clustering algorithms becomes the most suitable one for such scenarios. It has wide variety of applications such as document classification, delivery store optimization, identifying crime localities, customer segmentation, fantasy league start analysis, insurance fraud detection, rideshare data analysis, cyber-profiling criminals, call record detail analysis, automatic clustering of IT alerts and much more. K-means clustering algorithm remains as very effective and efficient clustering algorithm till date. In this project, we have implemented K-means clustering algorithm using python to cluster the rate of influence of diseases.

6. Experimental results

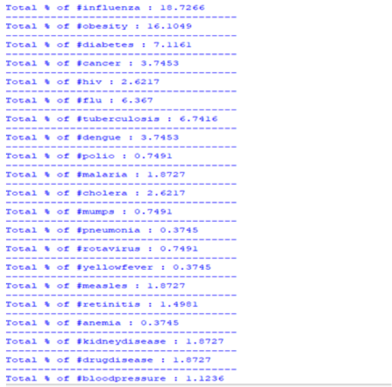
Once the diseases have been clustered using K-means algorithm using python, the rate of influence of diseases obtained are,



```
 RESTART: C:\Users\RAVINDRAN\AppData\Local\Programs\Python\Python37-32\code1.py
Occurrences of the word:
--------------------------------------------------
Distribution of various disease| Rate of influence
--------------------------------------------------
#influenza                         50
#obesity                           43
#diabetes                          19
#cancer                            10
#hiv                               7
#flu                               17
#tuberculosis                      18
#dengue                            10
#polio                             2
#malaria                           5
#cholera                           7
#mumps                             2
#pneumonia                         1
#rotavirus                         2
#yellowfever                       1
#measles                           5
#retinitis                         4
#anemia                            1
#kidneydisease                     5
#drugdisease                       5
#bloodpressure                     3
#colourblind                       4
#viralfever                        2
#smallpox                          8
#sars                              3
#ebola                             1
#chickenpox                        6
#stroke                            4
#eyedisease                        4
#anthrax                           2
--------------------------------------------------
```

Fig. 10. Count of diseases obtained from dataset.

Once we have collected the count, we need to obtain the rate of influence of diseases. From Fig11 we can infer the trending diseases in recent years. From the results, it is inferred that diseases have been classified in such a way that they are rated from high to low based on their rate-of-influence. According to the dataset taken by us, the diseases such as influenza, obesity and diabetes tops the chart. Polio has been reduced by 99% but not completely eradicated as it still remains in some countries and our result infers it very well.



```
Total % of #influenza : 18.7266
------------------------------------
Total % of #obesity : 16.1049
------------------------------------
Total % of #diabetes : 7.1161
------------------------------------
Total % of #cancer : 3.7453
------------------------------------
Total % of #hiv : 2.6217
------------------------------------
Total % of #flu : 6.367
------------------------------------
Total % of #tuberculosis : 6.7416
------------------------------------
Total % of #dengue : 3.7453
------------------------------------
Total % of #polio : 0.7491
------------------------------------
Total % of #malaria : 1.8727
------------------------------------
Total % of #cholera : 2.6217
------------------------------------
Total % of #mumps : 0.7491
------------------------------------
Total % of #pneumonia : 0.3745
------------------------------------
Total % of #rotavirus : 0.7491
------------------------------------
Total % of #yellowfever : 0.3745
------------------------------------
Total % of #measles : 1.8727
------------------------------------
Total % of #retinitis : 1.4981
------------------------------------
Total % of #anemia : 0.3745
------------------------------------
Total % of #kidneydisease : 1.8727
------------------------------------
Total % of #drugdisease : 1.8727
------------------------------------
Total % of #bloodpressure : 1.1236
```

Fig. 11. Rate-of-influence of diseases.

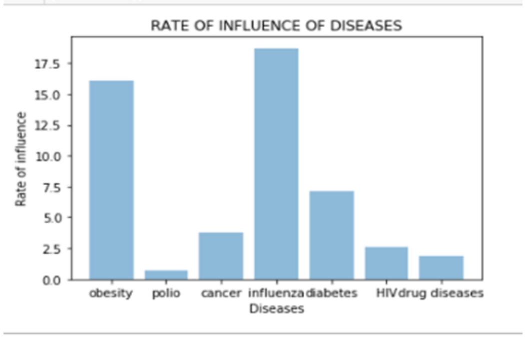


Fig. 12. Graphical Representation of Most-influenced Diseases.

Conclusion

The project emphasizes on observing public health on one of most prominent social media platform called Twitter. K-means clustering algorithm is used to sort-out the rate of influence of diseases that has been ruling the world in recent years and time-efficiency and accuracy has been achieved when working with large data-sets. The diseases listed under "Ruling-Diseases" list helps researchers to mine more valuable information and create awareness among public, conduct workshops and seminars to make people be aware of the most over-rated diseases in recent times.

References

[1] Sumit Sidana, Sihem Amer-Yahia, Marianne Clausel, Majdeddine Rebai , Son T. Mai , and Massih-Reza Amini "Health Monitoring on Social Media over Time" in IEEE Transactions On Knowledge And Data Engineering, Vol. 30, August 2018.

[2] L. Manikonda and M. D. Choudhury, "Modeling and understanding visual attributes of mental health disclosures in social media," in Proc. CHI Conf. Human Factors Comput. Syst., 2017, pp. 170–181.

[3] Michael J. Paul and Mark Dredze "You Are What You Tweet: Analyzing Twitter for Public Health", in Proc. Int. Conf. Weblogs social media, 2011.

[4] Thomas Davidson, Dana Warmsley, Michael Macy, Ingmar Weber, "Automated Hate Speech Detection and the Problem of Offensive Language", in Proc. Int. Conf. ICWSM, March 2017, pp.512-515.

[5] Yubao Zhang, Student Member, IEEE, XinRuan, Student member, "Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending ", Haining Wang, Senior Member, Hui Wang and Su He, IEEE Transactions On Information Forensics And Security, Vol. 12, No. 1, January 2017

[6] S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo, "Tweet4act: Using incident-specific profiles for classifying crisis-related messages," in 10th Proc. Int. Conf. Inform. Syst.Crisis Response Manag., 2013.

[7] L. Hemphill and A. J. Roback, "Tweet acts: How constituents lobby congress via twitter," in Proc. ACM Conf. Comput. Supported Cooperative Work Social Comput., 2014, pp. 1200–1210.

[8] U. Pavalanathan and M. De Choudhury, "Identity management and mental health discourse in social media," in Proc. 24th Int. Conf. World Wide Web, 2015, pp. 315–321.

[9] F. Bouillot, P. Poncelet, M. Roche, D. Ienco, E. Bigdeli, and S.Matwin, "French presidential elections: What are the most efficient measures for tweets?" in Proc. 1st Edition Workshop Politics Elections Data, 2012, pp. 23–30.

[10] L. Hemphill and A. J. Roback, "Tweet acts: How constituents lobby congress via twitter," in Proc. ACM Conf. Comput. Supported Cooperative Work Social Comput., 2014, pp. 1200–1210.

[11] A. Ceron, L. Curini, and S. M. Iacus, "Using sentiment analysis to monitor electoral campaigns: Method matters-evidence from the united states and italy," Social Sci. Comput. Rev., vol. 33, no. 1, pp. 3–20, 2015.

[12] P. Barber_a, "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter Data," Political Anal., vol. 23, no. 1, pp. 76–91, 2015.

[13] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 151–160.

[14] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," J. Amer. Soc. Inf. Sci. Technol., vol. 60, no. 11, pp. 2169–2188, 2009.

[15] S. Wang, M. J. Paul, and M. Dredze, "Exploring health topics in Chinese social media: An analysis of sinaweibo," in Proc. AAAI Workshop World Wide Web Public Health Intell., 2014.