Medinsights: Twitter based Platform for Health Care Analytics

Akansha Jain
Calpine Labs
UVJ Technologies pvt. ltd.
Cochin, India
aakanshaajainn@gmail.com

Sreejith Cherikkallil
Calpine Labs
UVJ Technologies pvt. ltd.
Cochin, India
sreejith.cherikkallil@calpinetech.com

***Abstract*—Twitter is a social media platform where the tweets convey opinions, but the interpretation of this unstructured data can be very time-consuming. Medinsights is a data science platform which leverages machine learning and natural language processing technologies on twitter data. This platform helps to analyze the inherent value of the data extracted from tweets such as diseases, treatments, symptoms linked to healthcare domain, which can only be fully exploited through deep data analytics. Medinsights takes particular health-related query on medicine, disease, brand, medical hashtags as input and converts corresponding twitter data into actionable information in order to gain insights and provide diverse recommendations. Medinsights analyses the sentiment, trend, prevalent location to provide meaningful inferences. Medinsights extracts and proposes associated symptoms, diseases, treatments, drugs and brands based on user query from tweets. For this study gradient boosting classifier is used for tweets classification into medical and non medical domain. Word2vec word embeddings with feed forward neural network is used for sentiment analysis. Conditional random field (CRF) is used to extract medical entities from tweets. Medinsights can be helpful to research correlation from tweets and analytics related to healthcare domain.**

***Index Terms*—CRF, FFNN, Gradient Boosting Classifier, Machine Learning, Natural Language Processing, Twitter, Word2vec.**

I. Introduction

Twitter currently has reached 336 million average active users every month [1]. It is the biggest live open data source for a variety of domain, some part of which has obvious clinical data for the healthcare industry. The value of data is grasped only when this raw information is converted into the knowledge that helps make decisions. Using data science strategy, healthcare organizations can benefit on increasing volumes of data and medical knowledge in an organized, strategic way. Also, individual clinicians can use that knowledge to improve the safety, quality, and efficiency of the care they provide. This project aims to analyze Twitter data and extract various informations like sentiment, prevalent location, trend, diseases, treatments, symptoms, etc. Also, it applies various natural language processing & machine learning technologies such as classification, entity recognition, etc. to gain insights into data related with healthcare domain.

Medinsights is a Twitter based healthcare analytics platform having a web-based user interface to access its various functionalities. It is a medical domain specific engine to fetch live tweets, performing sentiment analysis on tweets, finding the prevalence of query in different locations, analyzing its trend, creating a word cloud of related words with user's query, etc. Another important aspect of Medinsights is to extract relevant information like symptoms, disease, treatment from tweets and present all information as summary called inference. The aim of creating such platform is to help general public as well as professionals in the medical domain to gain not just knowledge, but real information that comes from individualś experiences, shared openly using micro-blogging sites like Twitter. The project has three sections: Medinsights Analytics, Medinsights Recommender and Medinsights Inference, refer Fig. 1 to see the full structure.

A. Medinsights Analytics

This part of the project implements different types of data analysis on Twitter data extracted related to userś query. Each of the analysis is graphically represented for the better understanding and simplification of the insight. The first step to any type of analysis is to collect the appropriate data, and for this study, the data is straightly taken live from Twitter. The user queries and the tweets having the keywords that user mentioned are extracted. This is done using Twitter's API.

1) Classification: The presence of medical keywords in tweets does not necessarily signify that people have the illness, or are using the pharmacological substances (drugs), or are exhibiting certain symptoms [2]. Not all tweets extracted are based on the context (medical) of healthcare domain. Many times they are used otherwise, For example, *"The most dangerous heart disease:: Sharp memory.."*, this tweet mentions medical entity like "heart" and "disease", but is not relevant in terms of context. Hence, before doing any sort of analysis, the first step is to classify tweets into relevant and non-relevant category. It is also required to pre-process the query as well as the tweets retrieved before giving them to classifier. The result which is a collection of relevant tweets is fed to different analytics modules in the project as mentioned below.

2) Sentiment Analysis: This module gives the overall sentiment of the Twitter users on the entered query. Sentiment analysis is a good way to understand the emotions and attitude of people around the world. The sentiment here is positive or negative. This would be helpful in knowing how the world is reacting towards a particular disease, drug or any other health-

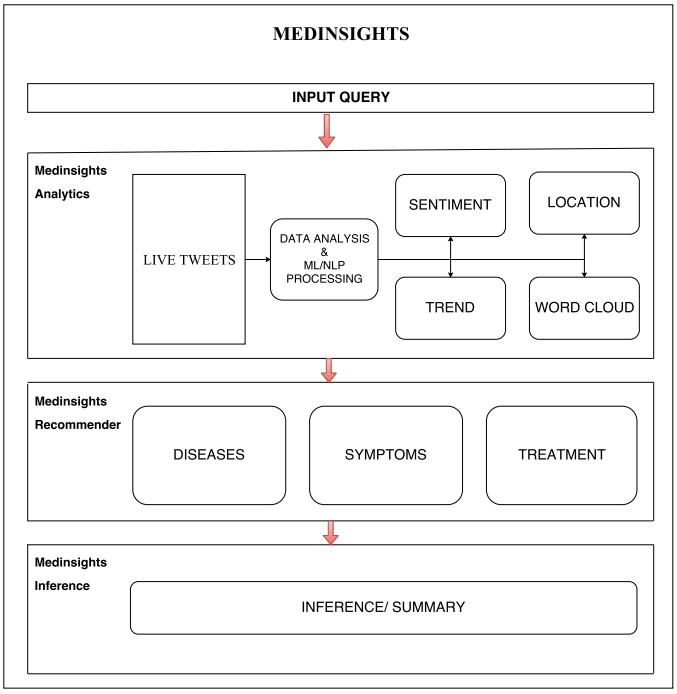

Fig. 1. Overview of Medinsights's Architecture.

related query. It has been a vastly covered topic in the field of research on Twitter data.

3) Location Analysis: A concerned user always wonders if the particular disease is affecting others, or where else some particular symptoms is breaking out. It would be so much better if there was some tool to inform about the same. This module does exactly and informs about how prevalent the users query is with respect to the location, i.e where in the world it is been talked about. Also, the intensity of the prevalence is depicted using a graphical representation in which colored dots mark the Geo-location on the world map.

4) Trend Analysis: Twitter Analytics, the service from Twitter, tells about user's tweets performance which is a great tool for analyzing their activity on Twitter. This service is limited to private use and can't be used to find trend about some query or hashtag. This module tells time trend of the entered query, which means how much a particular topic (users query) is being talked about over a time period. The trend is calculated per hour and per day basis. This work targets past one week's tweets and finds out how the queried entity varied in popularity with time. The trend is plotted graphically in the manner of a time series graph.

5) Word Cloud: The current world is flooded with information. The ability to express views instantly using social media platform like Twitter can be utilized to find intra information relations. This module gives information about what are the other entities, that people all around the world are talking about. The extracted top hashtags/keywords are beautifully represented as a word cloud where the most important or highest rank word is the biggest in size and likewise the size of the word is decreased on the basis of relevance.

B. Medinsights Recommender

Recommender systems are everywhere now, whether browsing through a movie or music website, shopping online, socializing on Facebook or Instagram. Phrases like, "*You may also like this*", "*You may know this person add them as your friend*", are often seen. This is possible because of artificial intelligence that learns from user's past behavior about what he may be interested in. In the same way, Medinsights Recommender is used for recommending information about diseases, treatments (drugs), symptoms, news based on the query. It learns from the data on that query and the user is suggested solutions and extra knowledge on the healthcare domain, derived from data generated by other users on Twitter which means these recommendations are real time query based.

C. Medinsights Inference

A query always demands a result, and this project gives the same in different analyses and recommendations. The inference is something which is logically derived from the knowledge that is acquired through experience. On the basis of above analysis, this section of the project will produce inferences as summary. These insights will help the user to make decisions.

The remainder of this paper is organized as follows. Literature survey is discussed in Section II. Section III explains the System Design. Experiment Setup and Results are discussed in section IV. The conclusion and future works are stated in section V.

II. Literature Survey

The following section explains past researches about identifying use of Twitter data in various fields of analytics, recommender and inference systems. Also, about existing state of the art methodologies used by various researchers.

Twitter has become useful social networking platforms for the general public to share thoughts, ideas and opinions. The real-time nature of the social platform and people being very open and candid about their health issues in tweets increases the authenticity of the data. Several studies have considered using Twitter for tracking various trends, including political opinions [9], sentiment [10], news tracking [7], earthquake monitoring [8]. Vance et al. [6] examined the advantages and disadvantages of utilizing web-based social networking to spread general wellbeing data in youthful grown-ups. Pros incorporate ease and fast transmission, while cons included absence of source reference and introduction of assessment as truth [6].

The focus is to use machine learning and natural language processing techniques to achieve this projectś goals. One such supervised machine learning technique called SVM classifier was used by Michael Paul et al. [4] to classify billion of tweets among health related and others classes. Support vector machine is widely used supervised machine learning algorithm for classifying the tweets in categories health or non health

related. The first section of this project in Medinsights Analytics implements that. The performance of ML classifiers need improvements.

Past research has explored Twitter for applying sentiment analysis on the various trending news and social topics to study the impact on the masses. Genes et al. [10] analyzed New York Storm response messages of Twitter users, C Adrover et al. [11] identified adverse effects of HIV drug treatment & associated sentiments using Twitter. These cases are studied by various researches since people are very responsive about them on social media, specially on Twitter. All these experiments throw the light on how important and available it is to understand people's opinions and attitude towards any crucial topic using Twitter. Research showed, Word2vec tool is the current state-of-the-art in generating distributed word representations from natural language data [14]. Medinsights Analytics utilizes value of Word2vec to include semantics while finding sentiment of tweets.

Twitter has been a great source for estimating trends [5] and finding location [12], [13] where certain disease outbreak happened, or where the certain medicine worked. For example, Wagner et al. [5] concluded that the vaccine program actually reduced the influenza like illness rates by 14% in england. Such studies are called Sentinel Surveillance, which collects health statistics for chronic diseases or risk factors, often based on geography. Syndromic Surveillance tracks trends in medical conditions over time as described by Paul et al. in [4] & [3]. These observations helped to work on Trend and Location section in Medinsights Analytics part of this project. The word cloud is also seen as a great tool to describe the results and trending keywords.

For the Medinsights Recommender and Medinsights Inference, the main objective is to find the medical entities (diseases, symptoms, treatment, drug) from the Twitter data. Paul et al. [3] associated symptoms, treatments and general words with diseases (ailments). They also proved that their Ailment Topic Aspect Model (ATAM) performed better than LDA which produced some topics related to diseases, but most did not clearly indicate specific ailments. In previous studies, Twitter has been used to extract information mentioned in tweets and then finding correlations with official surveillance data [2]. Topic analysis of widely used medicinal drugs [2] using LDA and SVM has been done too. Research showed, n-gram language models [14],[17], and other techniques are deployed to find this useful information, with respect to medical domain. Linear-chain CRF as used by Yepes et al. [15] is a popular algorithm to extract named entities from the text. CliNER [16] based on CRF and SVM is used for named entity recognition in clinical text of electronic health records, which is used in Medinsights Recommender of this project.

However, researchers are still striving to improve the performance of the ML Algorithms. These approaches mainly used n-gram bag-of-words or characters as ML features. These features suffer from the curse of dimensionality because the total dimension of each tweet's text is equal to the vocabulary size, which can overfit the models. Additionally, these features do not consider semantic relations between words, which can result in poor performances in some cases [2].

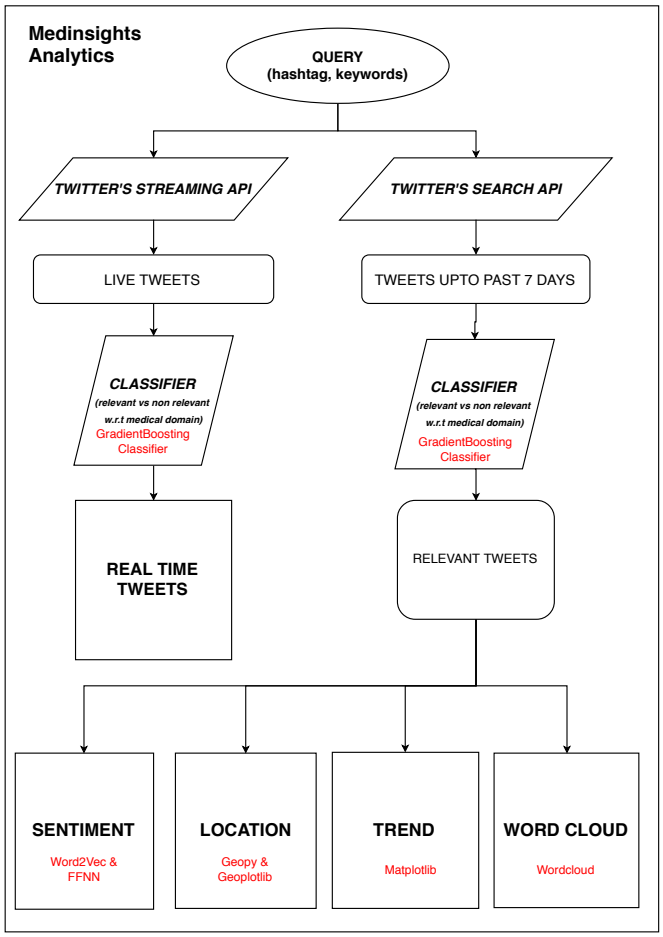

Fig. 2. Medinsights Analytics Work Flow Diagram.

III. System Design

The following section explains how the full platform is designed, refer Fig. 1, its work-flow, and the technology implemented. The Medinsights Analytics section cleans the user's query, collects the tweets relevant to medical domain. Then, displays the time trend, word cloud, sentiment and location based prevalence of the query based Twitter data. Medinsights Recommender suggests user about the symptoms, drugs, treatments, tweets and news linked with the query. All these information is forwarded to Medinsights Inference section, which gives user an overall summary that helps in making smart decisions.

A. Dataset Preparation

Dataset was manually prepared for training the ML model classifying tweets as relevant and non relevant with respect to medical domain context. The dataset consists of a thousand labeled tweets. The label is encoded as "0" for non relevant and "1" for relevant, refer Table I for sample dataset. Data was acquired using Twitter's API [23]. The API was given

TABLE I
SAMPLE RELEVANT (1) VS NON RELEVANT (0) TWEETS

Tweet	Label
"America suffers from gun disease"	0
"I have Alzheimer's disease when I read Shiaś tweets."	0
"One good thing about music, when it hits you, you feel no pain"	0
"Cure for #nausea? I can't wait to try this out http://bit.ly/2tnFdys . Hold an alcohol prep pad 2.5 cm from nose and inhale deeply for up to 60 sec. better than 4 mg zofran "	1
"Chronic bronchitis is characterized by the presence of a productive cough that lasts for 3 months or more per year for at least 2 years."	1
"#Lupus neurological symptoms: depression, seizures, cognitive dysfunction, mood disorder, cerebrovascular disease, polyneuropathy, anxiety"	1

relevant & non relevant keywords as input. Relevant keywords include symptom, disease or drug name for ex, *"Zofran", "Fever", "Bronchitis"*. Non relevant keywords include ambiguous words related with medical domain, for ex, *"sick", "tired"*, etc. This data is then parsed and cleaned by removing retweets (RT), URLs and emoticons.

B. Medinsights Analytics

The proposed system design is shown in Fig. 2 which describes the flow of the analytics process. The first step is to take the input query and live tweets are extracted using Twitter's Streaming API and tweets up to past 7 days are taken from Twitter's Search API. These tweets are fed to a supervised machine learning classifier, trained with our prepared dataset.

1) Classification: For creating a relevant vs non relevant classifier, we experimented with different algorithms. Out of which Gradient boosting (GB) performed well and gave an overall accuracy of 96%. GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n_classes_ regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification, which is our case, is a special case where only a single regression tree is induced.

2) Sentiment Analysis: For Sentiment Analysis, 3 layered FFNN (feed forward neural network) is used for classification of tweet in positive or negative sentiment. The word embeddings or word vectors prepared using Word2vec[24] are fed as input to the sentiment classifier. The classifier is trained with 1.6 million labeled tweets taken from a source [18]. Word2vec is a popular deep learning based word embedding method to consider the semantic of the text. On the other hand, neural nets are more advanced when considering machine learning models for classification and they certainly perform better. The activation function used in first two layers is 'relu' and for last layer is 'sigmoid'. FFNN generally learns using gradient descent non linear optimization technique. Each tweet is analyzed and classified to either positive or negative category. The overall sentiment of the query is calculated and then graphically shown as output.

3) Location Analysis: Location Analysis requires location data of tweets. The coordinates of tweetś are first searched, if not found, then the place of the tweet is searched, if not found, then the userś location from its profile is taken assuming the tweet was posted from there. Geo-Coding is required which converts a location to latitudes & longitudes. For this task, Geopy [19] library with Nominatim API is used to connect with OpenStreatMap to get coordinates. Once coordinates are acquired, they are plotted using library called Geoplotlib [20].

4) Trend Analysis: Trend Analysis tells about the frequency of query over a course of time period. The timestamps of relevant tweets is collected and grouped by per hour per day basis. The frequency of tweets is plotted as a time trend. The maximum of past 7 days trend is possible to plot. For this purpose, matplotlib [21] library is used.

5) Word Cloud: Word Cloud takes relevant tweets that are pre processed, cleaning out the non english words, stop words, emoticons, urls, etc. The processed tweets are given to wordcloud [22] package. This calculates the most frequent words appeared in the tweets related with userś query. These words, along with the size depicting intensity are colorfully plotted.

C. Medinsights Recommender

Symptom-Disease-Treatment (SDT) extractor based on CRF sequence modeling algorithm is trained to extract medical entities from collection of Twitter data. CRF also known as Conditional random field, which belongs to class of statistical modeling, is used in pattern recognition, machine learning for structured prediction. CRF is undirected probabilistic graphical method often used in POS tagging, named entity recognition, sequence labeling tasks for natural language processing. This project uses CliNER, a CRF based entity recognizer for medical domain.

D. Medinsights Inference

The outputs from the analytics and recommender sections are summarized in inference section. This gives user an insight from all of the information gathered from Twitter data.

IV. EXPERIMENTAL SETUP AND RESULTS

This section describes the comparison of different algorithms and methods that were used and their results. Also, discusses the output of all the modules of Medinsights.

A. Classification

Machine Learning models for classification such as Decision trees, Naive bayes, Support vector machines, & Ensemble models like Random forest, Gradient boosting, Stochastic gradient descent were tested on our dataset. We used 80% of our data for training and 20% for testing. It was concluded that Gradient boosting performed best with 96% accuracy out of all on the given data. Table II shows the comparative results.

TABLE II
COMPARISON OF DIFFERENT ML CLASSIFIERS

Classifier	Accuracy %
RandomForest	0.90
Decision trees	0.91
NaiveBayes	0.88
SVM	0.87
StochasticGradientDescent	0.93
GradientBoosting	0.96

TABLE III
RESULT OF SENTIMENT ANALYSIS

Query	Positive Tweets %	Negative Tweets %
#chronicillness	66.66	33.34
#dengue	22.22	77.78
#lungcancer	50.0	50.0
#lupus	42.85	57.15
#pneumonia	80.0	20.0

B. Medinsights Analytics

Following are the results shown for user query on #dengue.

1) Sentiment Analysis: The trained neural network was able to achieve 86% of accuracy on the training and 76% on validation test. Sentiment analysis showed that only 22% of user's are positively talking about dengue, rest 78% were negative. It can be inferred that dengue is spreading and people aren't happy with the current scenario. Table III shows overall sentiments for more user queries.

2) Location Analysis: Location analysis shows where in the world #dengue is spreading. It was found that India, US and countries from European continent were mainly seen to be affected. The results are shown in Fig. 3.

3) Trend Analysis: Trend analysis is performed for #dengue considering past seven days tweets. Observation was as shown in Fig. 4. The graph throws light on the possibility of outbreak of dengue during 14-03-2018 to 15-03-2018.

Fig. 3. Location Analysis of query #dengue.

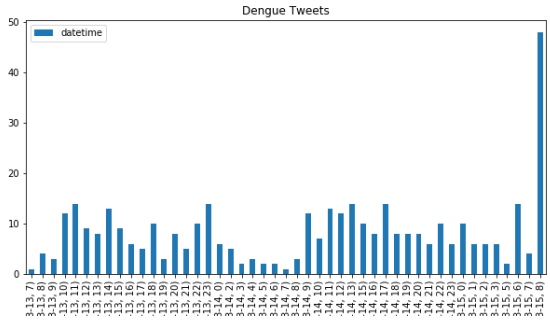

Fig. 4. Trend Analysis of query #dengue (13-03-2018 to 15-03-2018).

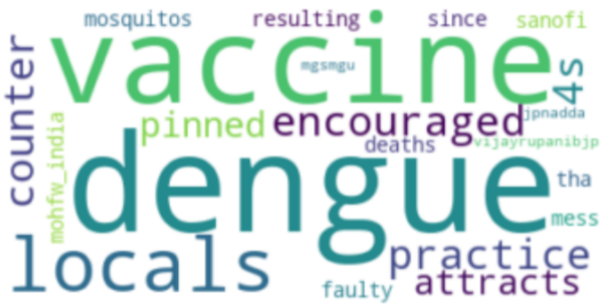

Fig. 5. Word Cloud of query #dengue.

4) Word Cloud: Giving relevant tweets on #dengue to this module, we can supposedly infer from keywords that "locals" are "encouraged" to get dengue "vaccine" since it is causing "death" and people are tweeting to "practice" measures to stay safe probably. Fig. 5 shows output of this module.

TABLE IV
INFORMATION EXTRACTED FROM TWEETS

Tweet	Problem	Test	Treatment
"Usually it is not mosquitoes that carry #dengue virus from one island to another, It is the people that carry virus for the mosquitoes."	"mosquito", "virus"	NA	NA
"#Dengue is a viral disease, transmitted by the infective bite of a particular mosquito known as Aedes Aegypti."	"a viral disease"	NA	NA
"US dengue expert says Sanofi ignored warning, no blood tests made for Dengvaxia"	"dengue", "dengvaxia"	"blood tests"	NA

C. Medinsights Recommender

As of now, we used CRF based CliNER tool to find out medical entities, such as "problem", "test", "treatment" from a given tweet. Table IV shows some sample tweets on #dengue and extracted medical information.

D. Medinsights Inference

The resulting inference template will look like the text box shown. All the results from each module in analytics and recommender are given to this template as input. The output gives summary of query, where bold and italics words are input from other modules. See the result of inference for #dengue.

Medinsights Inference

You entered ***#dengue***. It seems that ***23%*** users all around the world are reacting positively and ***77%*** are reacting negatively about it. The ***#dengue*** appears to be prevalent in ***India***, ***US*** and some ***European countries***. ***#dengue*** usage seems to be popular during time period : ***14-03-18 to 15-03-18***. The dominant keywords related with ***dengue*** are ***dengue, vaccine, locals, encouraged, practice***. The tweets on ***#dengue*** indicates ***mosquito, dengvaxia, virus disease*** as possible root cause and ***blood tests*** as remedial measure.

V. Conclusion and Future Scope

Medinsights presented classification method for identifying relevant tweets with respect to medical domain with accuracy above 96%. The different analysis like sentiment analysis, trend, location and word cloud helped to gain insights in popularity of the query. Given CliNER based approach to extract medical named entities such as a Problem, Test and Treatment from collection of tweets showed promising results. The inference system is helpful in deciding further actions and increases the knowledge base of our users on their given query.

The future scope of Medinsights includes (i) Deep learning models for classification, (ii) Finding out correlation between tweets and public data sources, (iii) Collaborative approaches to provide better recommendations. The application can be extended to support: (i) City level monitoring system to track disease outbreaks, (ii) User review based system to find drug side effects, (iii) Review mechanism on queried topics.

Acknowledgment

This work was fully funded by Calpine Labs, UVJ Technologies, Kochi, India. Also, this research was remarkably supported by Mr. Sreejith C, Data Scientist, Calpine Labs. We are also immensely grateful to Mr. Bijeesh Devassy, Project Manager, Calpine Labs and Dr. Asharaf S, Associate Professor, IIITM-K for sharing their pearls of wisdom with us during the course of this research. We also thank all of the colleagues from Calpine Labs, UVJ technologies, Kochi, India, who provided insight and expertise that greatly assisted the research.

References

[1] Statista, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
[2] Kagashe, Ireneus, Zhijun Yan, and Imran Suheryani. "Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data." *Journal of medical Internet research* 19.9 (2017).
[3] Paul, Michael J., and Mark Dredze. "A model for mining public health topics from Twitter." *Health* 11 (2012): 16-6.
[4] Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." *Icwsm* 20 (2011): 265-272.
[5] Wagner, Moritz, et al. "Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content." *Journal of medical Internet research* 19.12 (2017): e416.
[6] Vance, Karl, William Howe, and Robert P. Dellavalle. "Social internet sites as a source of public health information." *Dermatologic clinics* 27.2 (2009): 133-136.
[7] Lerman, Kristina, and Rumi Ghosh. "Information contagion: An empirical study of the spread of news on Digg and Twitter social networks." *Icwsm* 10 (2010): 90-97.
[8] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.
[9] Tumasjan, Andranik, et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." *Icwsm* 10.1 (2010): 178-185.
[10] Genes, Nicholas, Michael Chary, and Kevin Chason. "Analysis of Twitter users sharing of official New York storm response messages." *Medicine* 2.0 3.1 (2014).
[11] Adrover, Cosme, et al. "Identifying adverse effects of HIV drug treatment and associated sentiments using twitter." *JMIR public health and surveillance* 1.2 (2015). Mowery, Danielle, et al. "Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study." *Journal of medical Internet research* 19.2 (2017).
[12] Stefanidis, Anthony, et al. "Zika in Twitter: Temporal variations of locations, actors, and concepts." *JMIR public health and surveillance* 3.2 (2017).
[13] Kim, Annice E., et al. "Using Twitter data to gain insights into e-cigarette marketing and locations of use: an infoveillance study." *Journal of medical Internet research* 17.11 (2015).
[14] Sarker, Abeed, and Graciela Gonzalez. "A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities." *Data in brief* 10 (2017): 122-131.
[15] Yepes, Antonio Jimeno, Andrew MacKinlay, and Bo Han. "Investigating public health surveillance using twitter." *Proceedings of BioNLP 15* (2015): 164-170.
[16] W. Boag, K. Wacome, T. Naumann, A. Rumshisky. CliNER: A Lightweight Tool for Clinical Named Entity Recognition. (poster) AMIA Joint Summits on Clinical Research Informatics 2015. San Francisco, CA
[17] Sarker, Abeed, and Graciela Gonzalez. "Data, tools and resources for mining social media drug chatter." Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). 2016.
[18] Twitter Data, http://keenformatics.blogspot.in/2015/07/sentiment-analysis-lexicons-and-datasets.html
[19] Geopy, http://geopy.readthedocs.io/en/1.10.0/
[20] Geoplotlib, https://github.com/andrea-cuttone/geoplotlib
[21] Matplotlib, https://matplotlib.org
[22] Wordcloud, https://github.com/amueller/word_cloud
[23] Twitterś API, http://docs.tweepy.org/en/v3.6.0/api.html
[24] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).