Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019)
IEEE Xplore Part Number:CFP19OSV-ART; ISBN:978-1-7281-4365-1

Study on Machine learning based Social Media and Sentiment analysis for medical data applications

R. Meena[1], Dr. V. Thulasi Bai[2]
[1]Assistant Professor, Prathyusha Engineering College, Chennai, India
meenarajeswaran@gmail.com
[2]Professor, KCG College of Technology, Chennai, India
thulasi_bai@yahoo.com

Abstract: Due to the rapid advancements in social media, it generates voluminous data in almost different areas of applications. Large amount of potential health related data are being available in large scale in various sources of internet. We explored the small use case of social media data for a particular disease, cancer on three different social media platforms such as google trends, twitter and online forums with the sentiment analysis of the mined text. The study shows that people are more relied on social media for their health related queries and the twitter analysis shows that there is a significant raise in the percentage of positive sentiments in the tweets shared by the organizations and individuals on cancer.

Keywords: Social media, sentiment analysis, machine learning, polarity, text mining, subjectivity, opinion mining.

1 INTRODUCTION

Internet and social media are being a part of daily life. The number of people using the social media is ever - increasing day by day. By 2018, the number of users worldwide has reached 3.19 billion. Social media interaction generates voluminous data. The data if properly analyzed will yield valuable insights. Social media data analysis has certain ethical issues in using the data posted by the users. [1] The online health related data, even if it is available, there are challenges like lack of literacy in addressing the social health data, integrity of the information, extracting the region related health care data, detailed data on medications taken, identifying data related to particular illness, judging the data, etc., [2] Social networks along with other methodologies can be effectively integrated together to analyze and develop new web or mobile applications for health care management and diet / exercise recommendation models. [3] You tube videos can also be a source of creating awareness to the end user regarding a particular disease, which may reduce the anxiety disorder. [4] It is also possible to analyze the individual's social health environment using their social network connectivity. The assessment of risk of individual's possibility on developing health problems can be predicted by their relationship and connectivity in social networks. [5]

2 TOWARDS CANCER ON GOOGLE SEARCH TRENDS

Google search has become the most immediate relying source of information across the globe for almost any query where health care related queries also had their major role. The most common cancers such as breast cancer, bladder cancer, and colorectal cancer [6] were taken and their search histories are plotted using Google Trends.

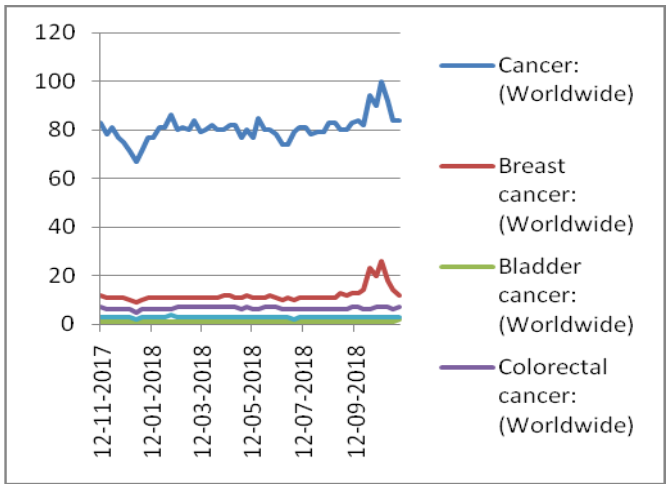

Fig.1. Some of the questioners posted on the Google search at the same period was "does coffee causes cancer", "coffee cancer", "melanoma cancer pictures", "peritoneal cancer", "early signs of lung cancer", "cancer vaccine", etc...,

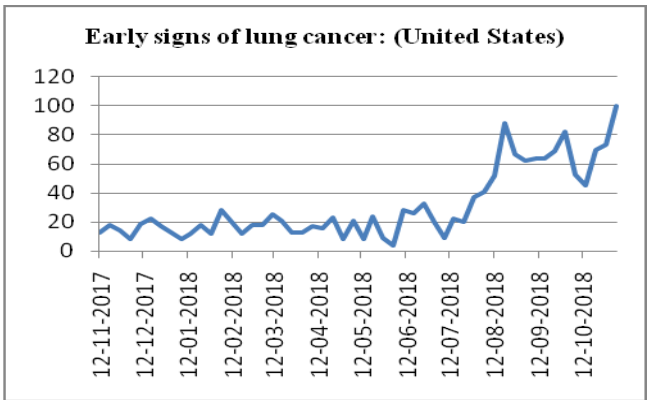

Fig.2. Graph showing the "Early signs of lung cancer" query posted on Google at United States" region for the period of one year

3 TWITTER FOR HEALTH CARE ANALYTICS

Twitter data can be used as the source of health care data analysis, which can be used by public health experts, clinical researchers, etc.., Lot of work, has been carried over on the tweets which gain great insights. Diabetes, diet, exercise, and obesity related tweets were extracted and linguistic analysis has been done to discover the common opinions and relationship among them. The strongest correlation among the topics was determined between exercise and obesity.[7] Using twitter based ads, 687 US based users of marijuana, a drug has been queried in an online survey and the pattern of usage along with the purpose of using the drug was identified. [8] ADR (Adverse Drug Effects) can be identified by negative sentiment analysis from online forums and tweets. The Health Improvement Network (THIN) database, General Practice Research Database (GPRD) which has a millions of patient records, tweets and other social networks contains user drug discussions was used to identify the ADR by Yellow Card Scheme[YR]. ADRMine uses sentiment analysis features to test the tweets in a number of domains using sentiment polarity and sentiment load using different lexicas. [9] Cancer, which is one of the most complex diseases, needs to be researched using data available from various aspects of cancer treatment. Twitter data, shared by many patients on their cancer treatments and their sentiments are available which, if utilized properly, can give significant insights. Patients tweet were analysed regarding brachytherapy for a particular period of time.[10] Online health forums can also be used to find the use of adverse health effects of banned drug usage by online surveys, tweets, face book and twitter ads etc.., Content analysis is also used to reveal the symptoms of mental health using depression related chats. [11][12][13][14] Tweets of a particular disease can be studied and the nature of tweets can be analysed for frequent queries and tweets. [15] Social networks can be used to track the discussion on vaccination opinions and eventually vaccination decision-making. [16] Using machine learning algorithms and statistical methods suicide-related terms were analysed in tweets thereby finding the depression and sucidality among the people. [17]

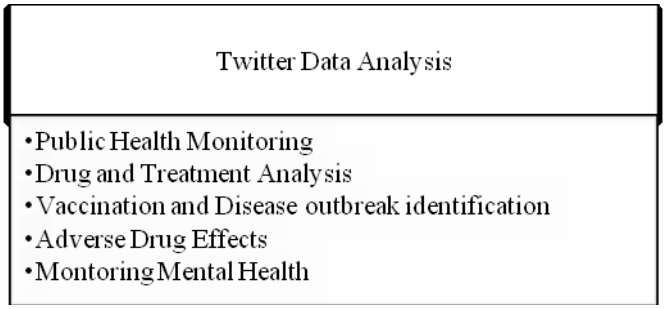

Fig.3. Applications of twitter data analysis

3.1 Sentiment Analysis

Sentiment analysis or opinion mining is defined as analysing the people's opinion and to gain the attribute of the text. Today we have data all over the internet in the form of tweets, face book comments, user comments on shopping sites, news articles, blogs, forums, etc..., The purchase decisions depend upon the positive, negative and mediocre comments and feedbacks given by the users. Sentiment analysis can also be applied to the healthcare industry for many applications such as, Feedback on the physician/drug, adverse drug effects, etc.., Linguistic analysis is needed to bring the insights of any sentence / feedback. Sentiment Analysis mainly has two main types. The Lexicon based approach follows tokenization followed by comparing the bag of words with the pre identified emotional words in the database. The overall score can be computed as positive, negative or neutral. The Machine learning based approach uses a train and test data set. The classifier can be trained with the training data set (eg: Classified tweets) and the test data will be given as input to it. It will give the desired result such as positive or negative.

Cancer Survivors Network - csn.cancer.org	FAERS FDA's Adverse Event Reporting System [pharmacovigilance]
https://connect.mayoclinic.org/	Med Effect Canada [pharmacovigilance]
Twitter data	Yahoo Answers
Medline, Medhelp, Daily Strength, AskaPatient	VigiBase[pharmacovigilance]
Breast Cancer.org	DrugBank[pharmacovigilance]
Cancer Genome Atlas	Breastcancer.org
THIN – Health Improvement Network	Komen.org

GPRD – General Practise Research Database	Bluelight.org
https://flusurvey.net/	Breast Cancer Support - csupport.org
dailystrength.org	healthboards.com
webmd.org	cancercompass.com
revolutionhealth.com	ehealthforum.com

Table.1. Sources for Health care Data Analytics

4 SAMPLE MODEL

Tweets related to cancer were extracted from teenagecancer, cancercounciloz, preventcancer, cancerschmancer, sylvestercancer, stanfordcancer which is different cancer foundation users available on twitter from various countries. The tweets were extracted using the python tweepy and twitters developer API. The key authentication has been provided to extract the twitter data from a particular user. The data has been extracted by using extractor object. We have extracted the tweets from nearly five twitter accounts which is related to cancer forums and the count of the tweets are 2000 of each user. A data frame has been created for the tweets. The metadata about every tweet such as the "created at, source, geo, re tweet count" can also be added to our data frame. Using numpy, a library in python the length of tweets, tweets having more number of likes can be computed.

The sample tweets retrieved

1000 recent tweets:

RT @User: Proud to have been a part of this novel drug's history in the phase I trial - and now excited by this data in lymphoma. My…
RT @User: Thanks, @bayareacancer 15th annual conference for making it a great closing session on making good choices. Look at all…
He learned she had breast cancer within days of giving birth. The Women's Cancer Center's care team and… https://t.co/zxmCFH4Opg
Join us for the Translational Oncology Program at # Annual Symposium! Panel discussions and lectures… https://t.co/fkfrbhW9RL
RT @User: Two scientists won the Nobel Prize in medicine for their work on cancer immunotherapy. Here's a visual explanation for how i…
RT @User : Steven Artandi, professor of medicine and of biochemistry at the School of Medicine, has been named the new director of @S…

The tweets retrieved were listed as samples (The names of the persons and organizations has been changed to "User" for privacy)

Tweets	Len	RT	SA
RT @User: Proud to have been a part of this novel....	140	2	1
Join us for the Translational Oncology Program....	140	0	0
Two scientists won the Nobel Prize in medicine for their...	140	1492	0
We need to invest in science...	140	4	1

Len – Length RT- Retweet SA- Sentiment Analysis

Suppose the tweets with more number of likes are considered as T_L and the tweets with more number of re tweets will be T_R. The relationship between the likes and re tweet has been obtained is such a way that

$$T_L \propto T_R$$

Sentiment analysis has been done on the extracted tweets from every user we have considered. The two step process of clean text and setting up the classifier for polarity analysis for the tweets has been followed. Python's text blob, a library for sentiment analysis was used for performing the polarity analysis. Polarity means the emotions expressed and subjectivity means the personal views and beliefs. The results obtained by the same was categorised into three classes such as Positive (P_O), Neutral (N_U) and Negative (N_E). The overall plot has shown that the positive and neutral tweets are more than the negative tweets. The result yields that most of the users are positively impressed and their mental health on the disease will be improved by the social network participation.

User	No. of Tweets extracted	No. of Tweets analysed
TeenageCancer	2000	1000
CancerCouncilOz	2000	1000
PreventCancer	2000	1000
cancerschmancer	2000	1000
SylvesterCancer	1000	1000
StanfordCancer	1000	1000

Table.2. User and Tweets

User	P_O	N_U	N_E
TeenageCancer	76%	21%	2%
CancerCouncilOz	55%	31%	13%
PreventCancer	60%	35%	4%
cancerschmancer	33%	64%	2%
SylvesterCancer	51%	41%	7%
StanfordCancer	44%	46%	9%

Table.3. Percentage of Positive, Neutral and Negative Tweets

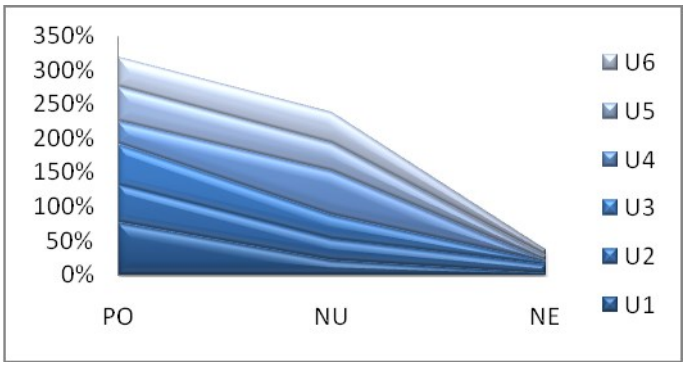

Fig.4. Graph showing the percentage of tweets

U1 - TeenageCancer, U2-CancerCouncilOz,
U3-PreventCancer, U4-cancerschmancer,
U5-SylvesterCancer, U6 – Stanford Cancer

Tweets Sentiment Visualization: Tweet visualization is more important to understand the significance of data using different visual contexts such as Timeline, Map, Heat map, Tag line, Affinity, etc.., A Sample visualization is presented using the 250 tweets retrieved from the particular date, which carries the key word "Breast Cancer".

Date	User	Tweet
11-12-2018 08:56	danagoulet	Hey @NFL instead of spending millions of dollars on outfits for the team, wherever it be breast cancer or the military etc... donate the money to the charity instead. You look like fools
11-12-2018 09:03	OncologyForum	RT @ELS_Oncology: #OncologyNews Immunotherapy Breaks Through in Breast Cancer https://t.co/QGcm2XnzYt
11-12-2018 09:04	evankirstel	RT @evankirstel: 'Artificial Intelligence can now help doctors😷 detect breast cancer.' #BreastCancerAwarenessMonth #Healthcare #IoT #AI #Cancer #DeepLearning #BreastCancer via @IrmaRaste @eViRaHealth #GlobalSummit18 https://t.co/lNyJTToXug

Table.4. Sample Tweets on the keyword "Breast Cancer"

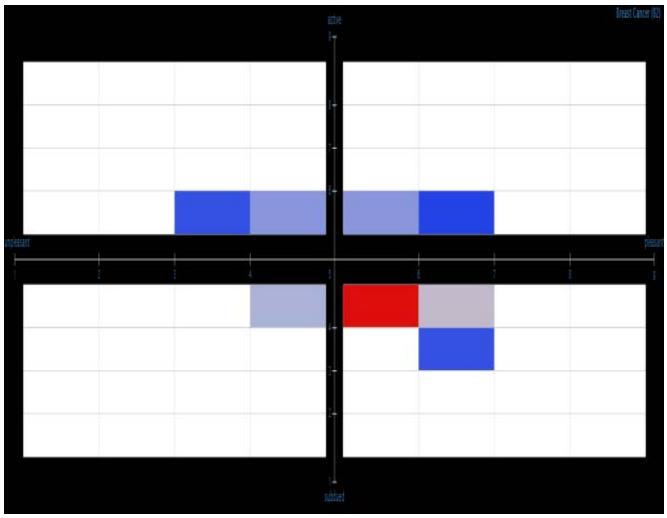

Fig.5. Heat Map on the tweets

4.1 Pattern Sentiment Analysis on Online forum posts

The discussion of online users on the topic "Squamous Cell Carcinoma of the Tonsil" has been retrieved from cancerresearchuk.org to perform the sentiment analysis using the two measures such as subjectivity and polarity.

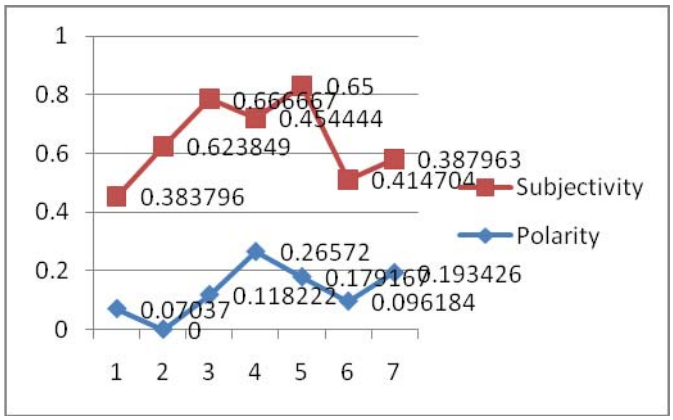

Fig.6. Graph showing the Subjectivity and Polarity of user comments

The average of the polarity from the user comments has been calculated as 0.153 and subjectivity as 0.5116.

5 CONCLUSION

Sentiment analysis and text based mining can be used as an ultimate tool for finding the user perception and public health intervention. The lexical and statistical analysis can act as a surveillance tool for health care data analysis. Specific methodologies can be applied to the data to obtain alleviate desired results.

REFERENCES

1. Golder S1, Ahmed S1, Norman G2, Booth A3. : Attitudes Toward the Ethics of Research Using Social Media: A Systematic Review, *US National Library of Medicine National Institutes of Health*
2. Atique S, Hosueh M, Fernandez-Luque L, Gabarron E, Wan M, Singh O, Traver Salcedo V, Li YJ, Shabbir SA: Lessons learnt from a MOOC about social media for digital health literacy, *Conf Proc IEEE Eng Med Biol* Soc. 2016 Aug
3. Ming-Hui Wen: Applying gamification and social network techniques to promote health activities, *2017 International Conference on Applied System Innovation (ICASI)*
4. Amy L.KotsenasMD[a] , MakalaArceBA[b], LeeAaseBS[b]Farris , K.TimimiMD[c], ColleenYoungBA[d]John T.WaldMD[a], : The Strategic Imperative for the Use of Social Media in Health Care , *Journal of the American College of Radiology,* Volume 15, Issue 1, Part B, January 2018, Pages 155-161
5. Amar Dhand [1,2], Charles C. White[3], Catherine Johnson[4], Zongqi Xia[5] & Philip L. De Jager [3,4]: A Scalable online tool for quantitative social network assessment reveals potentially modifiable social environment risks, *Nature Communications*, DOI: 10.1038/s41467-018-06408-6
6. https://www.cancer.gov/types/common-cancers
7. Amir Karamia, Alicia A. Dahlb, Gabrielle Turner-McGrievyb, Hadi Kharrazic, George Shaw Jr.a : Characterizing diabetes, diet, exercise, and obesity comments on Twitter*, International Journal of Information Management* 38 (2018) 1–6
8. Raminta Daniulaityte, Mussa Zatreh, Francois R.Lamy, Ramzi W. Nahhas, Silvia S. Martins, Amit Sheth,Robert G. Carlson : A Twitter-based survey on marijuana concentrate use, *https://doi.org/10.1016/j.drugalcdep.2018.02.033*
9. Ioannis Korkontzelos, Azadeh Nikfarjam, and Graciela H. Gonzalez, J : Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts , *Biomed Inform*. 2016 August
10. Thomas J[1], Prabhu AV[2], Heron DE[3], Beriwal S[4]: Twitter and brachytherapy: An analysis of "tweets" over six years by patients and health care professionals, *US National Library of Medicine National Institutes of Health*, doi: 10.1016 /j. brachy.2018.07.015.
11. Raminta Daniulaityte, Mussa Zatreh, Francois R. Lamy, Ramzi W. Nahhas, Silvia S. Martins, Amit Sheth, Robert G. Carlson : A Twitter-based survey on marijuana concentrate use, *https://doi.org/10.1016/j.drugalcdep.2018.02.033*
12. Loflin M1, Earleywine M2: A new method of cannabis ingestion: the dangers of dabs?,, doi: 10.1016/j.addbeh.2014.05.013. Epub 2014 May 28.
13. Daniulaityte R[1], Lamy FR[2], Barratt M[3], Nahhas RW[4], Martins SS[5], Boyer EW[6], Sheth A[7], Carlson RG[2]: Characterizing marijuana concentrate users: A web-based survey, *drugalcdep.2017.05.034.* Epub 2017 Jun 29.
14. Cavazos-Rehg PA[1], Krauss MJ[1], Sowles S[1], Connolly S[1], Rosas C[1], Bharadwaj M[1], Bierut LJ[1]: A content analysis of depression-related Tweets, *Comput Human Behav*. 2016 Jan 1;54:351-357
15. Sutton J[1], Vos SC[2], Olson MK[2], Woods C[3], Cohen E[4], Gibson CB[5], Phillips NE[5], Studts JL[6], Eberth JM[7], Butts CT[8]: Lung Cancer Messages on Twitter: Content Analysis and Evaluation.**,** doi: 10.1016/j.jacr.2017.09.043. Epub 2017 Nov 15.
16. Gema Bello-Orgaza Julio Hernandez- Castrob David Camachoa: Detecting discussion communities on vaccination in twitter, *Future Generation Computer Systems* Volume 66, January 2017, Pages 125-136
17. O'Dea a, Stephen Wan b, Philip J. Batterhamc, Alison L. Calear c, Cecile Paris b, Helen Christensen a : Detecting suicidality on Twitter, Bridianne, *Internet Interventions* 2 (2015) 183–188