ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 3, November 2017

# Intelligent Data Mining In Online Social

Joan Niveda J[1], Dr.G.Mariakalavathy[2] , M.Shalini[3]

PG Student, Dept. of CSE, St.Joseph's College of Engg., Chennai, Tamilnadu, India[1]

Professor, Dept. of CSE, St.Joseph's College of Engg., Chennai, Tamilnadu, India[2]

Assistant professor, Dept. of CSE, St.Joseph's College of Engg., Chennai, Tamilnadu, India[3]

**ABSTRACT**: Intelligently capturing knowledge from social media has recently attracted much interest from several fields such as Biomedical, Marketing and Advertising. Social media platforms like Twitter, Facebook and other online media network provides a platform for people to express their emotions in the digital world, which provide valuable information. People frequently ask their friends, relatives, and specialists for suggestion. By using this information several decisions can be made. Opinion can be collected from any individual about anything through review sites, blogs, and web forums. The proposed system includes a two-step analysis framework that focuses on positive and negative sentiment of treatment, in users' forum by analyzing their likes, posts and comments. It is used for the purpose of ascertaining user opinion of lung cancer treatment. Sentiment analysis on user forum posts from facebook is used to suggest the best treatment for lung cancer.

**KEYWORDS:** Social media, Sentiment analysis, Natural Language Processing.

## I.INTRODUCTION

In the past few years, there has been a huge growth in the use of social media sites such as Twitter. Due to the growth, companies and media organizations are increasingly seeking ways to mine facebook,Twitter for information about what people think and feel about their products and services.Social network has gained remarkable attention in the last decade. People are getting more interested in and relying on social network for information, news and opinion of other users.

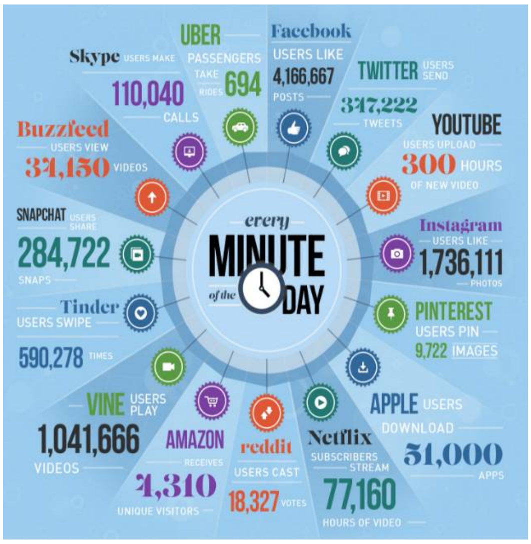


**Fig.1. Estimated Data Generated Every Minute on Social Media.**
http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/

Social networks[1] are important sources of online interactions and contents sharing, approaches evaluation influences observations feelings opinions and sentiments expressions that can be found out in text, reviews, blogs, discussions, news, remarks, reactions, or some other documents. Before the advent of social network, the homepages was popularly used in the late 1990s which made it possible for average internet users to share information. However, the things happening on social network now seem to have changed the *World* Wide Web (www) into its intended original creation. Social network platforms enable rapid information transactions between users independent of their location. Many organizations, individuals and even government of countries now follow the activities on social network [2]. The network enables big organizations, government official, celebrities and government bodies to obtain knowledge on how their audience reacts to postings that concerns them out of the enormous data generated on social network as shown in Fig.1. The network permits the effective collection of large-scale data which gives rise to major computational challenges. However, the application of efficient data mining techniques has made users to acquire valuable, useful and accurate knowledge from social network.

Data mining [3] provides a wide range of techniques for detecting useful knowledge from massive datasets. Data mining techniques are used for information retrieval, machine learning and statistical modeling. These techniques employ data pre-processing, data analysis, and data interpretation in the course of data analysis.

## II. RELATED WORKS

Data mining techniques have been found to be capable of handling the three dominant disputes with social network data namely; noise, size anddynamism. Some of the sentiment analysis techniques and an overview of tools used to analyse opinions conveyed on social network are discussed below.

### A. Opinion Analysis on Social Networking Sites

Users' opinions on social network [4] sites can be referred to as discovery and recognition of positive and negative expression. Various methods have been developed to analyze the opinion arising from products, services or events on social network. Some of the methods are discussed below.

### B. Aspect-Based Opinion Mining

Aspect-based also known as feature-based analysis is the process of mining the area of entity consumers has reviewed. This is because not all aspects/features of an entity are often reviewed by consumers. It is then necessary to summaries the aspects reviewed to determine the division of the overall review whether they are positive or negative. Sentiments expressed are simpler to analyze than the rest. According to [5]aspect-based opinion problem lies more inforum discussions andblogs than in product or service reviews. The aspect/entity (which may be a computer device) reviewed is either 'thumb up' or 'thumbdown', thumb up being positive opinion while thumb down means negative opinion.

### C. Homophily Clustering in Opinion Mining

Opinion of influencers over social network is based largely on their personal views and cannot be hold as absolute fact. However, their opinions could affect the decisions of other users on diverse subject matters. Opinions of influential users on Social network often count, resulting in opinion formation evolvement. Clustering technique could be used to model the formation of opinion by way of testing the affected nodes and unaffected nodes. Users who express the same opinion are linked under the same nodes and those with opposing opinion are linked in other nodes. This concept is referred to as homophily in social network [6].

### D. Opinion mining

Opinion mining or extraction is necessary in order to target the exact part of the document where the actual opinion is expressed. In opinion extraction, the more the number of people that give their opinion on a particular subject, the more important that portion might be worth extracting[5].

### E. Sentiment Analysis on Online - Social Network

Sentiment analysis can be referred to as discovery and recognition of positive or negative expression of opinion [7] by people on different subject matters of interest. An overview of some of the data mining tools used for sentiment analysis on social network is discussed below.

### F. Sentiment Orientation (SO)

Hand sellers make use of Sentiment Orientation (SO) for their rating standard in order to safeguard irrelevant or misleading reviews present to reviewers the 5-star scale rating with five signifying best rated while one signifies poor rating. In SO was used to improve the performance of mood classification. Live journal blog corpus dataset was used to perform evaluation of the above method. The experiment presented a modular proficient hierarchical classification technique which could be effectively implemented together with machine learning techniquesand SO attributes.

### G. Product Reviewsand Ratings

The reliance on the world wide web( social mediawebsites) for gathering information whilepicking choices about services or products has increased the needof research in the field of electronic-word-of-mouth. Products reviewingandratings often contains many sentimental expressions, a product can be rankeddepending on the mood of the user at the time.

### H. Reviews and Ratings Architecture

RnR (Reviews and Ratings)architecture produced complete description of a product and its service by using temporal dimension analysis with linear regressionand scatter plot. Tagging words was utilized for easily accessible domain ontology for the process of feature identification. Because tagging Parts of speechfor each word in entire reviews and also identifying opinion word could consume more time as well as cost more despite providing high accuracy.

### I. Aspect Rating Analysis Method

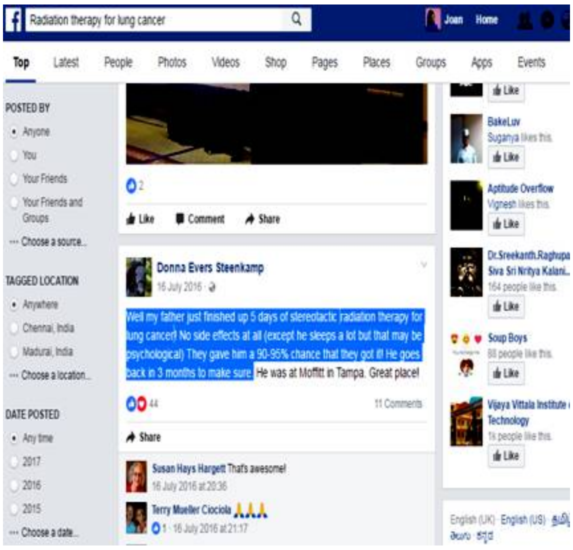
Aspect-rating is amathematical evaluation techniquewith respect to the aspect pointingto the degree of satisfaction depicted in the comments collected. Each aspect is captured using Probabilistic latent semantic analysis (pLSA). Itcan be utilized in place of formation of the phrase. The already existingpost is exploited to discover the aspect rating. Aspect cluster are nothing but words that stands for an aspect that reviewers are concerned andcomment about.Latent Aspect Rating Analysis (LARA) approach analyzes the user opinion by the process of text mining at the point ofrelated aspect.

### J. Sentiment Lexicon

Sentiment Lexiconis a collection of sentimental words which reviewers often make use in their expression. Sentiment lexicon is a list of most repeated words that would enhance the data mining techniques [8]while using mining sentiment in a document. Different kinds of sentiment lexicon could befor various subject matters. For example sentimental words that are used in sport are quite different compare to the words that are used in politics. Sentiment lexicon can be extended bymaking use of synonyms. Lexicon extension by the use of synonyms has a disadvantage of a word losing its meaning.

## III. PROPOSED SYSTEM

The proposed system suggests the best drugs and treatment methods based on users (doctor, patient and other related user) comments, likes and posts as shown in Fig.2. In this work comments has more priority than likes. It also identifies the users who can spread the information effectively.



Donna Evers Steenkamp
16 July 2016

Well my father just finished up 5 days of stereotactic radiation therapy for lung cancer! No side effects at all (except he sleeps a lot but that may be psychological) They gave him a 90-95% chance that they got it! He goes back in 3 months to make sure. He was at Moffitt in Tampa. Great place!

Susan Hays Hargett That's awesome!
16 July 2016 at 20:36

Terry Mueller Ciociola 🙏🙏🙏
1 · 16 July 2016 at 21:17

Fig.2. Online Social Media (Facebook)

The influenced user is ranked based on connectivity and his friend lists.The architecture of the proposed system is given below in the fig.3. The functions of the proposed system includes preprocessing data, Natural Language Processing (NLP), sentiment analysis, network based modeling.

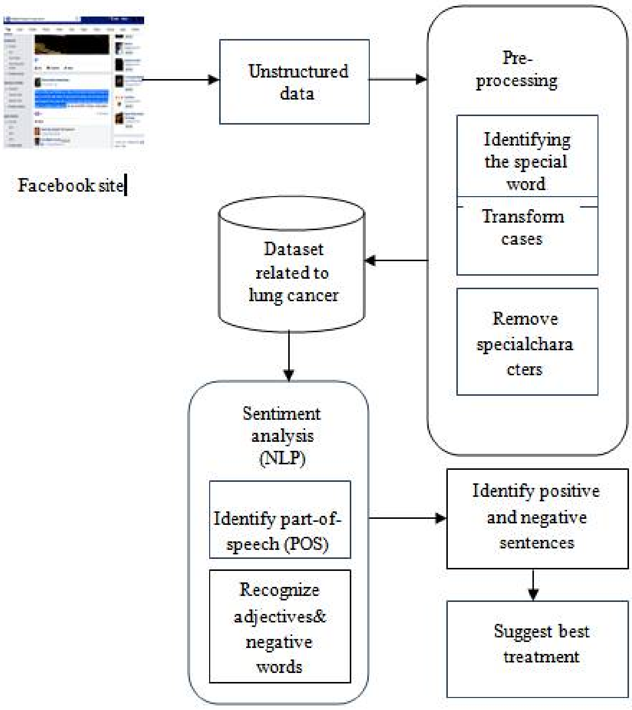

Facebook site
Unstructured data
Pre-processing
Identifying the special word
Transform cases
Remove specialcharacters
Dataset related to lung cancer
Sentiment analysis (NLP)
Identify part-of-speech (POS)
Recognize adjectives& negative words
Identify positive and negative sentences
Suggest best treatment

Fig.3. Architecture of the proposed system

A. Initial Data Search and Information Retrieval

The popular cancer message boards are captured in order to obtain the correct dataset that is very much related to cancer. Initially, the number of posts about lung cancer is considered. Lung cancer is chosen because, according to the work of previous researchers, it is the maximum observed cancer in the world. Some of thetreatments used by lung cancer patients are noted. Treatments such as radiation, cryotherapy and chemotherapy are taken into consideration and comments related to these treatments are analyzed as to which treatment is most welcome by the users (doctor, patient and other related user).

B. Preprocessing:

Preprocessing is an important task and a critical step in information retrieval (IR), Natural Language Processing (NLP) and Text mining In the field of Text Mining, data preprocessing is used for inferring factsand knowledgefrom rawdata.Information Retrieval (IR) is process of deciding which documents should be retrieved in order to satisfy a reviewer's demand for information.

Preprocessing consists of two steps namely tokenization and normalization process. The flow of preprocessing is shown in Fig.4. Emoticons and abbreviations are recognized as a result of tokenization process. For the normalization process, presence of Emoticons (for example, positive ':)' and negative ':(') in a message are omitted. And also Informal text such as all-caps (for example, I LIKE this show!!!) and character repetitions (for example, This treatment workss!! happyyyyyy")are identified. All-caps words are changed to lower case, and repeated letters are replaced by a single letter. Finally, any special comment tokens are(for example, user tags, and URLs) removed.

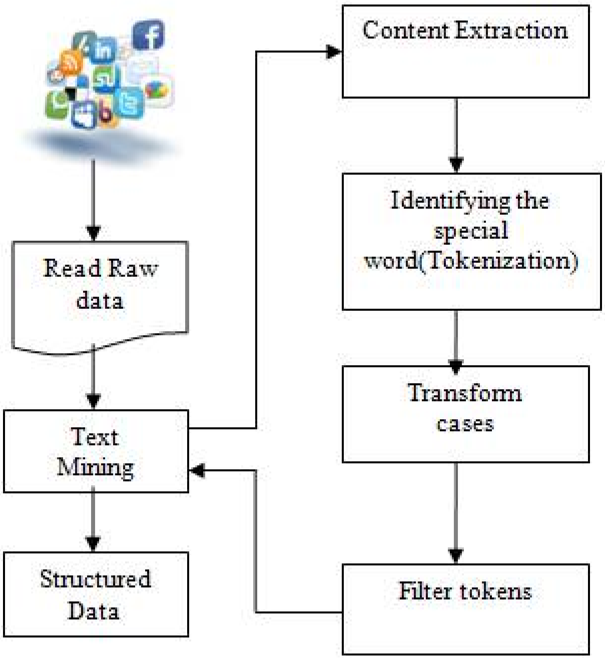


Fig.4.Pre-Processing

C. User opinion identification using NLP

Natural Language Processing (NLP)[10] approach consists of the following steps

1. Preprocessing: Read reviews, and then follow the below mentioned steps

a. Search part-of-speech (POS) and root for every word in the post.
b. Recognize adjectives in the text
c. Verify whether the neglected words comes before or after the adjectives.

2. Apply Procedures: Obtain attributes and link them with their adjectives that are mentioned in the first step.

a. Tag up to a maximum of two words highlighted by an adjective, stop when coming across aparticle, punctuation mark or verb.
b. Use the below procedures in order to form adjective phrases:

Adjective Phrase : < Attribute>< Adjective>
|< Attribute >< Neglect >
< Adjective >
Attribute : Simple Attribute(Positive)
|Compound Attribute(Negative)

c. Verify if the adjective phrase is already present in the adjectives table, determine its classification has either positive or negative or else classify and update.
d. Verify whether the adjective phrase is simple (positive) or compound (negative) in attributes table, else validate and update the attributes table.

D. Algorithm

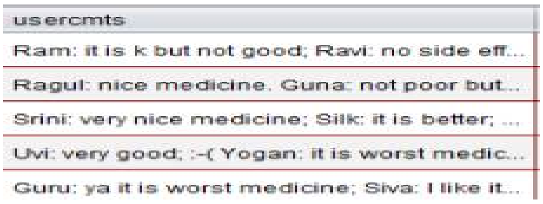
Sentiment Classification Algorithm is used to analyze particularwords or sentence combination to obtain a number which would tella user’s opinion from his messages.The algorithm utilizes list of sentimental keywords which containscollection of negative and positive words. AFINN is a list of English words rated for the integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Arup Nielsen in 2009-2011.

Very Negative (rating -5 or -4)
Negative (rating -3, -2, or -1)
Positive (rating 1, 2, or 3)
Very Positive (rating 4 or 5)

Categorize words as very negative to very positive and add some Medical-specific Adjectives.

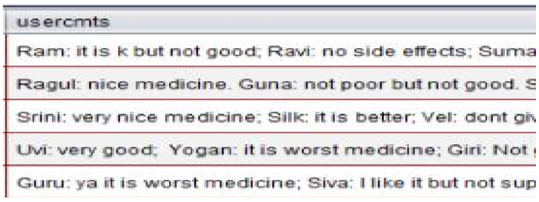
```
vNegTerms <- c(afinn_list$word[afinn_list$score==-5 | afinn_list$score==-4],”harmfull”,”very bad”)
nnegTerms <- c(afinn_list$word[afinn_list$score==-3 | afinn_list$score==-2 | afinn_list$score==-1], "very bad”,”harmful”,---------------)
posTerms <- c(afinn_list$word[afinn_list$score==3 | afinn_list$score==2 | afinn_list$score==1], "good",”great”, ---------------)
vPosTerms <- c(afinn_list$word[afinn_list$score==5 | afinn_list$score==4], "awesome")
```

## IV. RESULTS



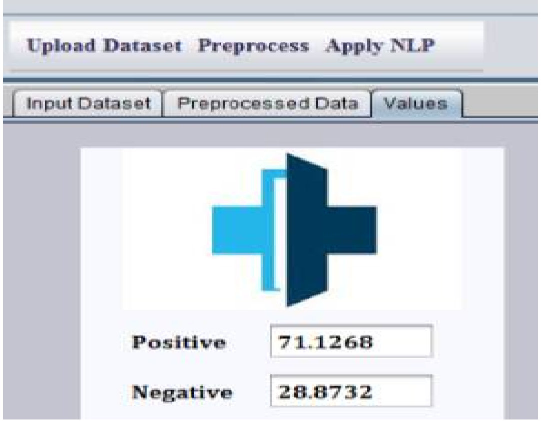
| usercmts |
|---|
| Ram: it is k but not good; Ravi: no side eff... |
| Ragul: nice medicine. Guna: not poor but... |
| Srini: very nice medicine; Silk: it is better; ... |
| Uvi: very good; :-( Yogan: it is worst medic... |
| Guru: ya it is worst medicine; Siva: I like it... |

Fig.5.Before pre-processing



| usercmts |
|---|
| Ram: it is k but not good; Ravi: no side effects; Suma |
| Ragul: nice medicine. Guna: not poor but not good. S |
| Srini: very nice medicine; Silk: it is better; Vel: dont gi |
| Uvi: very good; Yogan: it is worst medicine; Giri: Not |
| Guru: ya it is worst medicine; Siva: I like it but not sup |

Fig.6.After pre-processing



Positive 71.1268

Negative 28.8732

Fig.7.Identification positive and negative opinions using NLP

## V. CONCLUSION

In this paper, a novel method for inferring knowledge from the users' posts and comments on social networking sites such as facebook is proposed. Hereby, posts relevant to lung cancer are collected and the sentiment of the post is analyzed and finally the opinion of the user who posted is inferred. This sort of inference aids the patients to know the most current and most welcomed treatment. Future studies will require more up-to-date information for a clearer picture of user feedback on drugs and services.

Future Work will require more advanced identification of inter-socials dynamics and its impacton the members.such interests of research includeslikes of posts,rankings, , and friendships. Further works may also include analyzing posts that contains informal language terms such as slang.

## REFERENCES

[1] L. Dunbrack, "Pharma 2.0 – social media and pharmaceutical sales and marketing," in *Proc. Health Ind. Insights*, p. 7., 2010.

[2] S.Wasserman and K. Faust, Social Network Analysis: Methods and Applications. New York, NY, USA: Cambridge University Press, pp. 825, 1994.

[3] J. Hans and M. Kamber, "Data Mining: Concepts and Techniques". 2nd ed.Burlington, MassMA, USA: Morgan Kaufmann, 2006.

[4] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012

[5] Liu, B.: Sentiment analysis and opinion Mining. AAAI-2011, San Francisco, USA, 2011.

[6] McPherson, M., Smith-Lovin, L., Cook, J. M.: Birds of a feather: Homophily in social networks. Annual review of sociology, 415-444, 2001.

[7] BlessySelvam, S.Abirami, "A Survey on Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering , Vol. 2, Issue. 9, 2013.

[8] A. Akay, A. Dragomir, and B. E. Erlandsson, "A novel data-mining approach leveraging social media to monitor consumer opinion of sitagliptin," J. Biomed Health Inform. Vol: PP, Issue: 99.

[9] World Cancer Research Fund International. (2013, Dec.13).Cancer Statistics Worldwide. [Online]. Available:http://www.wcrf.org/cancer_statistics/world_cancer_statistics.php

[10] S. Bird, E. Klein, and E. Loper, Natural Language Processing With Python. Sebastopol, CA, USA: O'Reilly Media, pp. 504, 2009.

[11] Altug Akay, Andrei Dragomir, BjornErikErlandsson. "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care". IEEE journal of biomedical and health informatics, vol. 19, no. 1, 2015

[12] Pang, B. and L. Lee, Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of Conference on Empirical methods in natural Language Processing (EMNLP), Philadelphia, July 2002, 79 - 86. Association for Computational Linguistics, 2002.

[13] C. Corley, D. *Cook*, A. Mikler, and K. Singh, “Text and structural data mining of influenza mentions in web and social media,” *Int. J. Environ.Res. Public Health*, vol. 7, pp. 596–615, 2010.

[14] L.Toldo, “Text mining fundamentals for business analytics,” presented at the 11th Annual Text and Social Analytics Summit, Boston, MA, USA, 2013.

[15] SENTIMENT ANALYSIS:[Online].Available: //www.youtube.com/watch?v=kFGronMuchU