Mining User Generated Contents in Online Healthcare forum using Text Mining Techniques

M. SuryaPrabha[1], Dr.B.Sarojini Balakrishnan[2]
[1]Research Scholar
[2]Assistant Professor (SS),
Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.
[1]suryaprabhaphcs2015@gmail.com,[2]dr.b.sarojini@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: The development of online health care forum patients and health seekers can access the Internet and get health related information through their social media blogs and health care forums. These forum users are involved to post emotional based messages and health related posts. In this forum enormous amount of emotional based messages posted by different stakeholders that are patients, caregivers and healthcare experts. In this online healthcare forum patients /healthseekers post a question. Peers who reply to the questions asked that replies are based on more emotions oriented messages. For analyzing these messages to find the adoption based answers , Stakeholders and Topic identification in the messages and Sentiments (emotions) to the messages. For this Framework we use various text mining methods as preprocessing , NLP methods ,and Machine learning algorithms as SVM- RBF(radial Basis Function) Classification , K-medoid clustering method and Lexicon Data Analysis (LDA) is used .In this paper Medhelp online healthcare forum messages were taken up for analysis.

Keywords: Online Healthcare forum, Text mining, Natural Language Processing(NLP), Knowledge Adoption Model, Stakeholder and Topic Analysis, Sentiment Analysis.

1. Introduction

Online health forums providing opportunities for people living with health conditions to connect with those who have put forward health problems, experiences and support. These forums such as Medhelp, Patients like me, Patients.info can be accessed via the Internet and are available to help others who live with or who are affected by health issues. Patients may provide this knowledge or healthcare providers may assist.These forum posts are published to everyone who has internet access to information relating to health/sickness and even to asoprivate authentication messages. This health information continue to open up linking means within this evolving scope for patient/healthcare seekers. (Allen C, I and al Vassilev 2016).

Current data reveals that over 70 percent of people who use the internet have been using the website to find health records, emphasise the strong demand for medical information, information on pharmas and comprehensive information on their disease-related messages and commentaries in online healthcare forums. Healthcare services are a popular source of health information. (Franz Grinvald H, Grosberg D and 2016).

Registered participants share their medical encounters with illness in these fora to offer solutions to their health issues. This user-generated posts are often focused on positive and poor views[1]. Those seeking health information in the online world will receive commercial services that deliver immediate messages to registered individuals [2,3] guidance from healthcare specialists, information dependent on patients' feelings, and even other random information. This material needs to be continuously analysed in order to evaluate usage when huge quantities of health material are checked. [4].

However, most of the Internet communications are not verified. Most messages and remarks published by users are not helpful or pertinent material. The vast quantity of content provided by users in health fora will mislead the seekers of health-related facts. Thus, challenges for patients and information providers are which determinants impact user's information acceptance and how they detect the most useful information for their needs. The development of an understanding of the actions of patients in online healthcare records will allow the online medical community to design and process and improve the efficiency of online health information.

The med help forum is considered for this research study is a forum has abundant source of medical, health and wellness information. In this research work, the discussion boards of 3 women related diseases were considered. The message contents of miscarriages, ovarian cyst and breast cancer are analyzed. These messages are to be analyzed by the text mining techniques and content analysis methods. In this work knowledge adoption of messages will be analyzed, then identify the stakeholders of the forums and the topics that were discussed more in each forum. The percentage of contribution of the stakeholders and the percentage of their contributions to

topics such as symptoms, complications, examinations, drugs, and treatments.This research framework will be helpful in determining the adoption decision for answers in given messages, stakeholders involved ,topics identify and sentiment expressions were considered in online healthcare forums to get valuable informationfor stakeholders, medical experts and healthseekers.

2.Literature study

Web-based forums on healthcare provide a medium for articulate communication through healthcare blogs between patients, nurses, doctors and other healthcare professionals to conduct and share medical knowledge, debates, recommendations and views[5]. Health forums are providing more information about disease related problems and emotional messages created by users[6]. In recent case studies the prevalence of various forms of medical knowledge has taken into consideration medicinal messages on blogs by mining the online communications sent by group members.In most trials, chronic diseases as Parkinson's disease and prevalent high mortality diseases including lung cancer, diabetes were consumed.

Online wellness forums are a medium for sharing stories, seeking out resources and improving their understanding of health with patients in a secure atmosphere. Members of wellness forums have been found to benefit from the online activities, leading to a better understanding of their symptoms and increased support for health with an effectiveness comparable to non-internet treatments. In conjunction with this, medical forums will provide valuable social promotion, inspiration and emotional support related to health. (Lewis LK, Maher CA, Ferrar K et al 2014).

Previous findings have found that prevention and diagnosis, recovery, care and long-term side effects of therapy are among the most prevalent conditions in which patients are involved. These studies investigate certain challenges in handling vast volumes of text material. Recent research used text mining techniques to explore the subjects of online health knowledge using this medical text provided by users.

Sentiment analysis is now being used for a number of online contexts such as healthcare apps for online shopping and so on. These methods of feeling analysis can be categorised into lexicon-based approach and a machine learning approach. Lexicon-based approach is dependent on lexicon-like emotion, a list of feeling-like phrases. A master's method which uses language features to study the words or characteristics of the given text.An integrated methodology incorporates lexicon and machine-based techniques and modeling. Denecke focuses on medical weblogs and classifies subjects of medical weblogs into two kinds: informational and affective. In order to classify the relevant dimensions of online examinations of healthcare practitioners, Brody used a text description based on Latent Dirichlet Allotment(LDA) models. [7]

Approach to classifying online research was suggested by author[8]. In this sentence review, a complex dictionary was made of predefined positive and negative terms in order to better find the feeling polarity of the words.

A approach for sentimental analysis at document level was suggested by the author [9] to identify the polarity of an entire news item. A dataset of news papers has been explored. On the whole news storey, text preprocessing was completed. In this stage, this score for the feeling of the article was given. The papers were classified as positive, negative or neutral.

In order to determine positive and negative language, the writer [11-12] has used python packages for classification of terms and SentiWordNet dictionary to explore opinion. In the news stories it can be measured the total impact of the good and bad feelings. This approach has been used to examine the effect of news headlines. This article considers the study of sentiments dependent on the sentimental wordnet approach based on Lexicon. We take Medhelp online healthcare forum for our experimental work. Text preprocessing was done using POS tagging, Text Tokenization and removal of unwanted text in the data. Senti-wordnet dictionary is to analyze the words of sentiment polarity based on positive, negative and Subjective to the messages.

3. Research Methodology

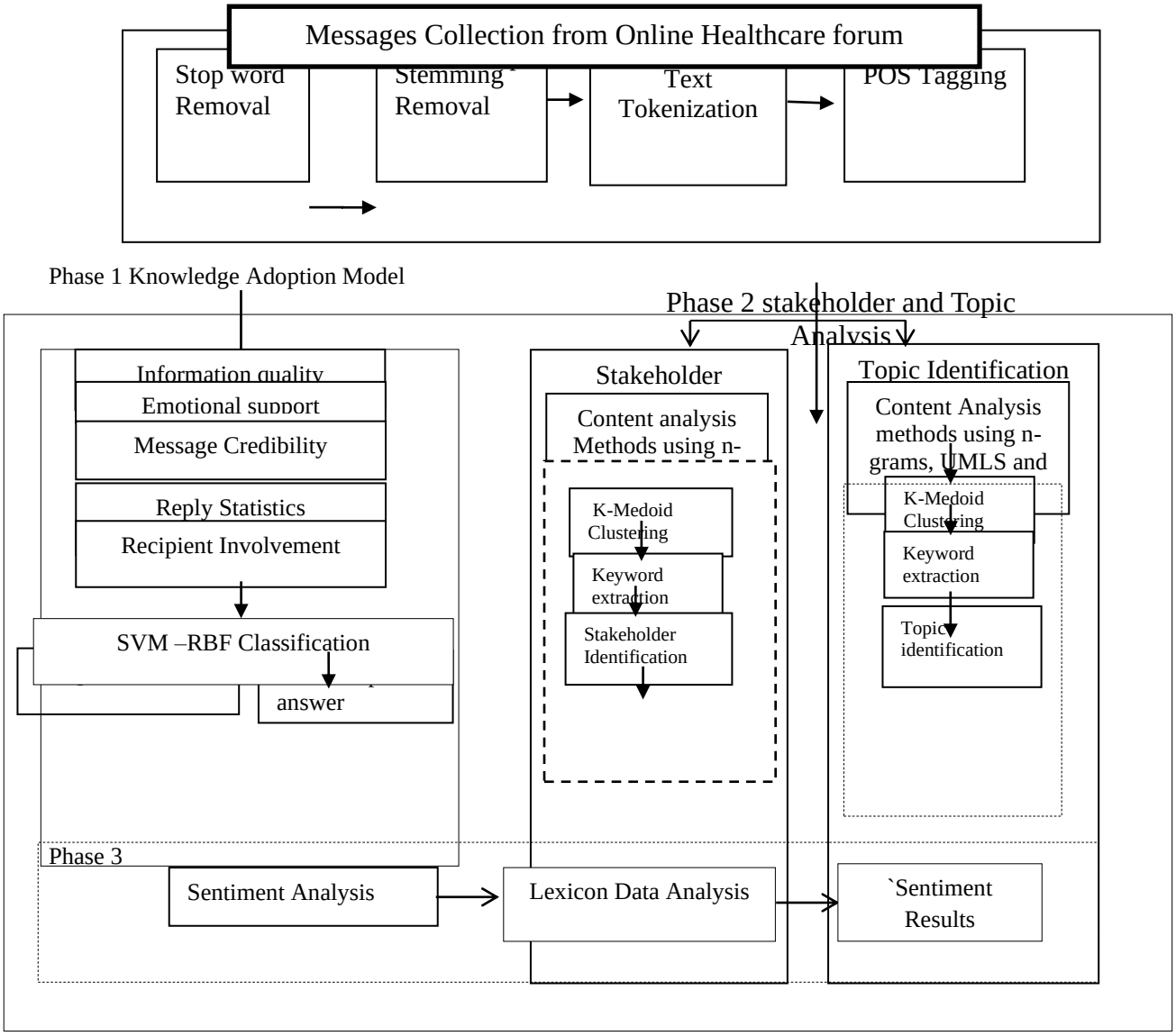

Figure1. Overall Research Framework

3.1 Dataset Description

Online healthcare forum MedHelp.org data is retrieved. This forum has more than 250 panels of various diseases related to fitness. The study has included three women's-related groups.

Table1. Dataset statistics Medhelp.org

Dataset name	Total messages
Breast cancer	480
Ovarian cyst	380
Miscarriages	410

3.2 Text-mining

In recent years, the techniques of Text Mining (TM) have been increasing due to the huge amount of text data generated in various forms like social networks, medical blogs, patient reports, health insurers and news outlets. The development and evaluation process involves vast quantities of unstructured text information assisted by software and methods to classify patterns, subjects, keywords and other attributes in the data. The process of text material is organized, semi-structuring and unstructured, such as text records, videos and pictures. Web blog contents are created by the user.It usually covers a wide range of similar subjects and algorithms for text analyses that include different groups, including data acquisition, the processing of natural languages and the mining of data.

3.3 Preprocessing Using NLP Methods

NLP is one of Text Mining's essential tools for text processing. The text is recognizable from the text interpreted by the computer. Computers and computers deal with tabular details or table-topics, excelling So that knowledge people chat, write or posts comments on online blogs is unstructured. NLP reflects the natural human language, such as voice and text. The NLP Text Pre Processing System (NLP) is a sub area of artificial intelligence in which its expertise includes computer-human experiences. These non-structured knowledge are anlyze on software to process NLP (Pratik Shukla, Roberto Iriondo 2021).

3.4 Stop word removal

Pre-processing is called the method of transforming data into correct computer understanding. The primary preprocessing approach is the filtering of unusable records. Unused words, text (data), are defined as stop words in the natural language processing.

Stop Word: Stop wording is a typical term in the text used to be deleted into text or documents (like "the," "a", "an," "in" etc.).

3.5 Stemming

The meaning in the stem (root) of the derived words in the text or the sentence is found by stemming. There is no further correct information available on these terms. (The ovaries). (example). In this case, ovary requires facts, but xenophobic words are taken away by stemming mechanism.

3.6 Part of Speech (POS)

The text mining tool is also part-of-the-spoken (POS) marking. In natural language processing technique that applies to classification of terms in a text (corpus), in correspondence with a specific section of the vocabulary, according to word meaning and context.Speech sections describe how a term is used in a sentence or paper, whether the word is a verb, a noun, an adjective, etc.

4. Knowledge Adoption ModelFor health care forum

Knowledge adoption Model framework deals with patients or users in healthcare forum to get better knowledge about their health related issues and disease related information.Thisinformation wereextracted from forum messages. These messages could be analyze by various Text mining approaches such as NLP preprocessing to preprocess the text and then to determine the knowledge adoption by analyzing the contents of the answers and their relevance to the question posted. The assessment is done based on five factors such as Information Quality, Emotional Support, Message Credibility, Recipient Involvement and Reply Statistics. These results could be analyzedby SVM-RBF(Radial Basis Function) Classification to provide the results. This work was carried out and results are provided in my previous work paper [14].

5. Stakeholders and Topic Identification Framework

The objective of this phase is to identify the different stakeholders and the topics that is widely discussed in the forum.

Pre-processed keywords must be classified as content-specific features, domain features and feeling-specific features. The n-gram study of content analysis, medical terms and the terminology of the relationship was used in order to classify stakeholders.N-gram analysis is used to extract the features. In this analysis keywords are extracted using n-gram techniques as unigram words, bigram words, trigram words and 4-gram words. Medical terminologies in the text were correlated with the Unified Medical Language System (UMLS). Kinship terms were mapped using UMLS semantic type such as family group, to identify family members related terms to identify the stakeholders[patients, Caregivers and Specialists] keywords and Topics[symptoms, Complications, Treatments, Examinations and Drugs] keywords in the messages posted in the online healthcare forum.This work was carried out and results are provided in my previous paper [15].

6. Sentiment Analysis

The study of feelings is one of the most effective approaches to interpret the views of persons or patients, their feelings, assessments and attitudes against them and their facets, as conveyed in a written text included in the web

blogs. In these terms, the Lexicon Based approaches can analysis. [4] Senti-wordnet dictionary is usually a Lexical Opinion Mine Source, described by a Senti-wordnet dictionary as the relations between words and opinions.The Lexicon-based sentiment words are categorized as positive and negative thought in approximately 6 800 English phrases. This collection was also begun with a series of names and adjective words, which are good or bad, and was extended by a process of discovery focused on semantic synonym and antonymic relations. SentiWordNet as well as lexicon methods include English words and feeling principles databases. The emotional based lexicon is used to identify the polarity of terms in the given messages which can be positive, negative and subjective. In the messages given, positive polarity defines positive terms and negative polarity defines negative words. The polarity of each word determines the sense of the phrase.

6.1 Sentiment Analysis Using Lexicon Based approach

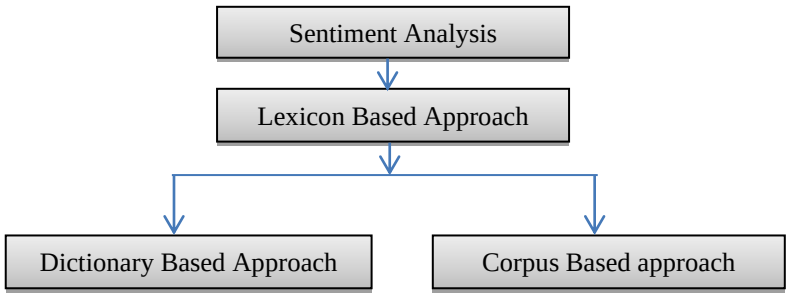

Figure2. Lexicon Based Framework

Sentiment analyses based on a lexical approach would be performed. The dictionary method for text messages is used in this research. In order to locate a constructive, negativist polarity in the statement or texts, the dictionary technique is used. The polarity of each word which combine the words of feeling of the sentence or the text. Via absolute terms of polarity of the words or phrases in each sentence the Polarity of a sentence is then determined.

6.2 Senti-wordnet

Senti-wordnet dictionary comprises information about the emotions or polarity expressed by words, phrases, or perceptions. In these dictionary usually provides one or more scores for each word. We can analyze the overall sentiment of an input sentence based on individual words. Senti-wordnet is essentially a lexical resource of emotion strength for opinion mining, with the relation between words and opinions centered on the Wordnet dictionary established.

6.3 Sentiment Polarity Calculation

Sentiment polarity categorizes the words as positive polarity, Negative polarity and subjective. Positive polarity defines the positive words in the given messages .Negative polarity defines negative words in the given messages. Subjective sentences generally refer to personal opinion.

$$\text{Positive (p)} = \frac{number\ of\ positive\ words\ in\ sentence}{total\ number\ of\ sentence} \quad (1)$$

$$\text{Negative (p)} = \frac{number\ of\ negative\ words\ in\ sentence}{total\ number\ of\ sentence} \quad (2)$$

Here the positive (p) denotes positive polarity of the words in given text and Negative (p) refers negative polarity of the given words.

$$\text{Emotion score} = \frac{Total\ number\ of\ p(w)+n(w)}{Total\ number\ of\ messages} \quad (3)$$

Here p(w) and n(w) defines the total number of positive words and negative words in the messages.

$$\text{Subjective polarity} = \frac{Subjective\ Polarity\ Words}{Total\ number\ of\ messages} \quad (4)$$

6.4 Results and Discussion

Table 1 Sentiment polarity results for positive, Negative and Subjective words.

Dataset name	Total messages	Positive words	Negative words	Subjective words
Breast cancer	750	16	62	24
Ovarian cyst	640	29	43	26
Miscarriages	590	19	57	23

6.5 Results

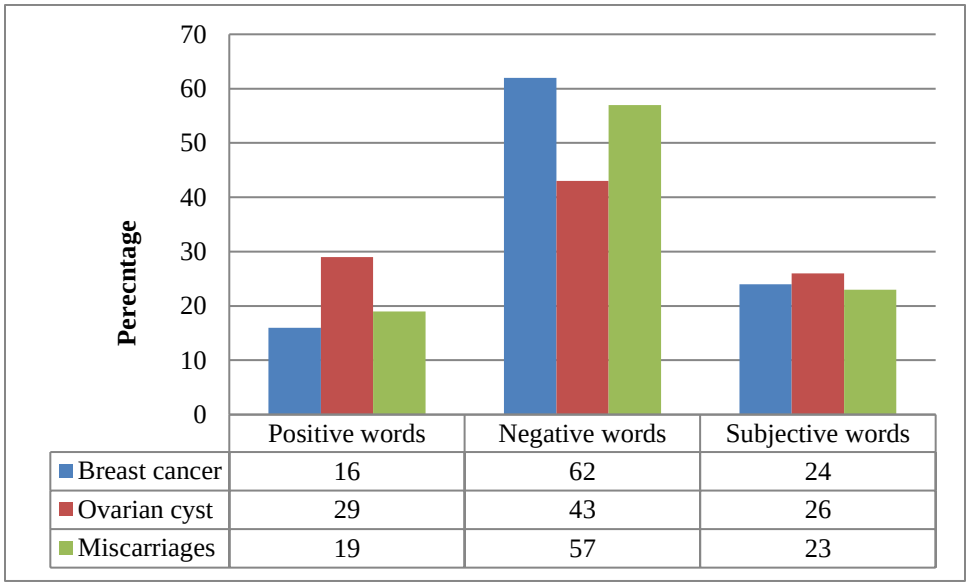

Figure 3.Results for Positive,Negative,Subjective polarity

In this results define the polarity of the positive words was higher in ovarian cyst data as (29%) than the Breast cancer and Miscarriages dataset. Then the Negative polarity of the words discussed high in the Breast Cancer data as (62% .Thus the subjective polarities of the words are higher in the ovarian cyst data as (26%).

Table 2: Percentage of Sentiments

Dataset name	Total Messages	Subjective words	Emotional Words
Breast cancer	750	76	22
Ovarian cyst	640	82	18

Miscarriages	590	73	25

6.6 Results

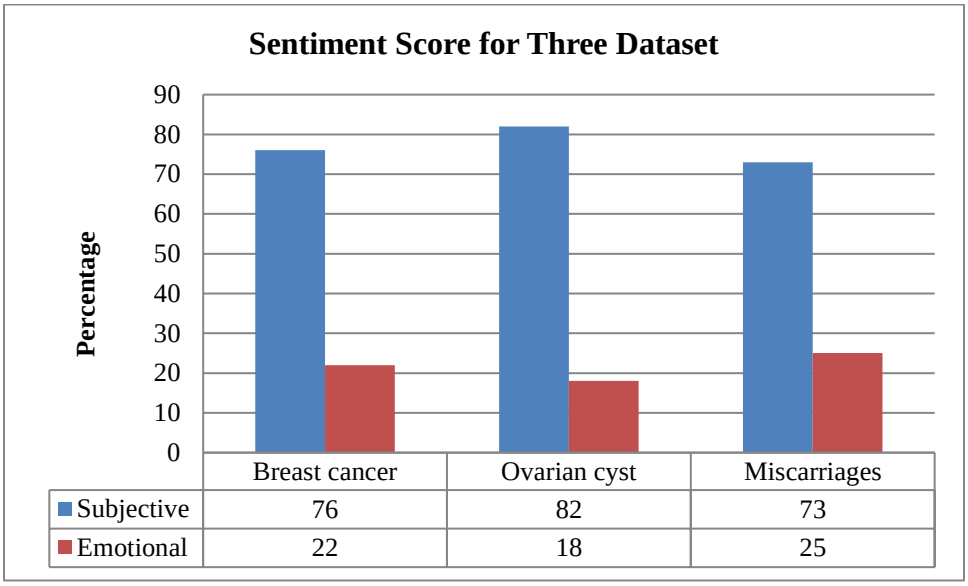

Figure.4 Results for Sentiment Score for Three Dataset

7. Conclusion

This research work has important suggestions on the field practitioners and patients /Healthseekers of online communities. This framework is to provide better helpfulness and health related Knowledge of user-generated contents of similar health or disease related feedback provided by the community participants. The users/health seekers in health care forum can be benefited from the framework by easily accessing helpful knowledge embedded in the text documents in online communities. In our First Framework was carried out by many factors to analyze the user generated contents by various text mining approaches and classification method to find the valuable Answers .In Second Framework is used for analyzing keywords of the messages based on posts .These Keywords are analyze by various content analysis method and clustering method to identify the users [patients, caregivers, Specialists] and medical related topics could be identified in the forum.In our final framework consist of analyzing the Sentiments[Emotions] based messages are analyzed and to find the Positive, Negativeand Subjective posts in the health care forum.

REFERENCES

1. Chuang KY, Yang CC", A study of informational support exchanges in medhelp alcoholism community. In: Yang SJ, Greenberg AM, Endsley M (eds)Proceedings of the 5th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP'12). Springer-Verlag, Berlin, Heidelberg,(2012),pp9– 17.

2. Im EO, Chee W An online forum as a qualitative research method: practical issues. Nurs Res 55(4):267–273.
3. Jin J, Yan X, Li Y, Li Y, How users adopt healthcare information: an empirical study of an Online Q&a Community. Int J Med Inform, (2016), 86:91–103
4. Monnier J, Laken M, Carter CL Patient and caregiver interest in internet-based cancer services. Cancer Pract(2002) ,10(6)pp:305–310
5. Xiao N, Sharman R, Rao HR, UpadhyayaSFactors influencing online health information search: an empirical analysis of a national cancer-related survey. Decis Support Syst (2014) 57:pp- 417– 427.
6. Rolls K, Hansen M, Jackson D, ElliottDHow health care professionals use social media to create virtual communities: an integrative review. J Med Internet Res, (2016),18(6):e166.
7. LinhVu ,Thanh Le, Ph.D. A lexicon-based method for Sentiment Analysis using social network data, Int'l Conf. Information and Knowledge Engineering, 2017, ISBN: 1-60132-463-4.
8. Ohana, B. and Tierney, B.,Sentiment classification of reviews using SentiWordNet. In 9th.it& conference, 2009, October (p. 13).
9. M. U. Islam, F. B. Ashraf, A. I. Abir and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary",20th International Conference of Computer and Information Technology (ICCIT), Dhaka, 2017, pp. 1-5..
10. Sentiment Analysis from News Articles,"International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-4.
11. A. Agarwal, V. Sharma, G. Sikka and R. Dhir, "Opinion mining of news headlines using SentiWordNet," Symposium on Colossal Data Analysis and Networking (CDAN), Indore, 2016, pp.1-5.doi: 10.1109/CDAN.2016.7570949
12. Qiu B, Zhao K, Mitra P, Wu D, Caragea C, Yen J, et al. Get online support, feel better: sentiment analysisdynamics in an online cancer survivor community. 2011 Presented at: 3rd IEEE International Conference on Social Computing..
13. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. 'When 'Bad' is 'Good": identifying personal communication and sentiment in drug-related tweets. JMIR Public Health Surveill 2016 Oct 24;2(2):e162
14. Prabha, M.S., Sarojini, B. Online Healthcare Information Adoption Assessment Using Text Mining Techniques. Mobile Network Applications 24, 1160–1165 (2019).ISSN 1572-8153, https://doi.org/10.1007/s11036-019-01253-3.
15. M.SuryaPrabha ,Dr.B.Sarojini ,A Novel Framework for Analysing Online Healthcare Information with Text Mining Techniques, ISSN 00975-2366,DOI- https//doi.org/10.31838/ijpr/2021.13.02.236.
16. Asraf Yasmin, B., Latha, R., & Manikandan, R. (2019). Implementation of Affective Knowledge for any Geo Location Based on Emotional Intelligence using GPS. International Journal of Innovative Technology and Exploring Engineering, 8(11S), 764–769. https://doi.org/10.35940/ijitee.k1134.09811s19