Sentiment Analysis on Myocardial Infarction Using Tweets Data

[1]Dr. R. Hemalatha, [2]M.B.Monicka
[1,2]Dept. of Computer Science, Tiruppur Kumaran College for Women, Tirupur, TamilNadu, India

Abstract
In 2016, the survey reports that 1.7 Million people die of Myocardial Infarction (MI), due to less medication facilities, less prevention care and treatment planning is top most analysis of effective disease risk assessment, through this we have take prevention using sentiment analysis of recent advancements, the text analytics have opened up new potential of using the rich information of tweet analysis, to identify the relevant risk factors in MI. To tackle the MI risk factors tweet analysis gives more remedy and care factors by users, also this leads to decrease of MI in India. Our system plays a machine learning approach using sentiment analysis using tweet dataset. Nowadays people suffering from MI such as cardiac arrest, high blood pressure, congestive heart failure etc. Twitter is an excellent resource for the MI Patients since they connect people who have with similar conditions and experiences. It provides the knowledge sharing about MI, plays a vital role through Opinion Mining system.

Keywords
Opinion Mining, Myocardial Infarction, Machine learning.

I. Introduction

A. Sentiment Analysis (or) Opinion Mining

The sentiment analysis it is also know as opinion mining, a product opinion is collected and categorized through the sentiment analysis by build a system. Automatic opinion mining frequently uses machine learning to mine the text for sentiment, a kind of artificial intelligence (AI). The opinion, attitude, sentiments, emotions, views etc., are extracts automatically by the Natural Language Processing (NLP) technique, in right context and categorize these into various groups like positive, negative, neutral etc.

For this research domain the various terms are subjectively analysis, subjectively detection, appraisal extraction and review and sentiment mining. The Opinion Extraction and Sentiment Classification are the two significant tasks involved in opinion mining and sentiment analysis. Opinion Extraction: From free text mining the opinionated phrases, in appropriate context. Sentiment classification: Depends on sentiment orientation it categories the opinionated phrases. The two main reasons have brought some unpredicted negative urbanization influences creating modern citizens more affected from diseases in nowadays are accelerated urbanization and improving living standard.

Sentiment analysis for medicinal services manages the analysis of human services related issues recognized by the patients themselves. It brings the patients feelings into point of view to make arrangements and alterations that could straightforwardly address their health issues. Sentiment analysis is utilized with business items to incredible impact and has outgrown to other application regions. Viewpoint based investigation of social insurance, prescribe the administrations and medications as well as present their solid highlights for which they are favored. Machine learning strategies are utilized to dissect a huge number of survey archives and finish up them towards a productive and precise choice. The directed systems have high exactness yet are not extendable to obscure areas while unsupervised strategies have low precision. More work is focused to enhance the precision of the unsupervised methods as they are more functional in this season of data flooding.

1. Sentiment Analysis Techniques

(i). Machine Learning

Machine learning based Sentiment Analysis or arrangement should be possible in two different ways:

- Sentiment Analysis by utilizing directed machine learning strategies and
- Sentiment Analysis by utilizing unsupervised machine learning procedures.

(a). Supervised Machine Learning

In Supervised Machine learning procedures, two sorts of informational collections are required: preparing informational index and test informational collection. A programmed classifier takes in the grouping variables of the report from the preparation set and the exactness in order can be assessed utilizing the test set. Various machine learning calculations are accessible that can be utilized extremely well to characterize the records. The machine learning calculations like Support Vector Machine (SVM), Naive Bayes (NB) and greatest entropy (ME) are utilized effectively in numerous examination and they performed well in the feeling characterization.

(b). Unsupervised Machine Learning

Lexicon Based Method is an Unsupervised Learning approach since it does not require prior training data sets. It is a semantic orientation way to deal with opinion mining in which sentiment polarity of highlights show in the given record are controlled by contrasting these highlights and semantic lexicons. Semantic dictionary contains arrangements of words whose sentiment orientation is resolved as of now. It arranges the archive by conglomerating the sentiment orientation of all assessment words display in the record, reports with more positive word lexicons is characterized as positive document and the documents with more negative word lexicons is classified as negative document.

Hybrid Technique: A few researchers combined the supervised machine learning and lexicon based techniques jointly to enhance sentiment classification performance. They considered both general reason lexicon and domain specific lexicon for identifying polarity orientation of sentiment words and feed these lexicons into supervised learning algorithm, SVM. They found that general purpose lexicon performed very poor while domain specific lexicon performed very well.

The system classified the sentiment in two steps: First the classifier is trained to predict the aspects and In Next the classifier is trained to predict the sentiments related to the aspects collected in step. Their system yielded around 66.8% accuracy.

2. Lexical Information

Understanding human emotions (also called sentiment analysis) is

a hard job for a machine, for which the computational intelligence technique may offer enhanced results. Regularly, semantic articulations and additionally paralinguistic includes in talked dialects (e.g., pitch, loudness, tempo, etc.) reveal the sentiments or emotional states of individuals. Prior research studies have developed sentiment lexicons using a dictionary technique and a corpus technique.

II. Related Work

[1] A paper with the movement of web innovation and its improvement there is a tremendous volume of information accessible in the web for web clients and a colossal measure of information is made as well. Web has transformed into a phase for web getting the hang of, exchanging considerations and bestowing bits of knowledge. Individual to individual correspondence areas like Twitter, Face book, Google+ are quickly grabbing pervasiveness as they enable people to share and express their points of view about subjects, have talk with various systems, or post messages over the world. There has been number of work in the field of estimation examination of twitter data. This survey bases chiefly on feeling examination of twitter data which is helpful to assess the data in the tweets where determinations are to an extraordinary degree unstructured, heterogeneous and are either positive or negative, or impartial now and again. In this paper, we offer a review and relative examinations of existing strategies for opinion mining like machine learning and lexicon-based procedures, along with assessment measurements. Over the information streams in twitter, we offer research using distinctive machine learning calculations like Naive Bayes, Max Entropy, and SVM (Support Vector Machine). We have likewise discussed basic complications and uses of Sentiment Analysis on Twitter.

In [2] work is related to document level sentiment analysis on review of movie dataset. They applied various machine learning approaches i.e. (support vector machines Naive Bayes and maximum entropy classification). From the IMDB website the dataset is taken for this research. Either with the stars and some numerical value scored authors review are only chosen. Features they have taken are unigrams, unigrams bi-grams, bi-grams, unigrams POS, adjectives, top 2633 unigrams, unigrams position. Results got pretty good in comparison to the human created baselines. SVM works best as compare to other classifiers. To categorizing the review as recommended or not, an easy unsupervised learning algorithm is presented with the paper. The reviews are categorized by using the normal [16] Semantic orientation of the phrases in the evaluation that hold adjectives or adverbs. The paper works in the sentiment analysis under document level. The word and semantic orientation of phrase is evaluated by using the point wise Mutual Information (PMI). They took reviews from opinions for different domains (Movies, Automobiles, Travel Destinations and Banks). They got ranges from 84 % of accuracy for the review of automobile to 66% of accuracy for the reviews of movie.

In [3] Sentiment analysis is the main tool which is used in this thesis. In this chapter, an elaborated description of sentiment analysis technique has been provided. Sentiment analysis, called as opinion mining, denotes the method of extracting subjective data from a source which contains objective and subjective information as well as other materials by applying NLP, text examination and computational linguistics. Opinion mining is divided into three steps as follows. First, the input will be divided into two parts; afterward each part will be tested to realize if it covers any sentiment, meaning that each part will be examined to notice if it is subjective or objective. Second, the subjective sentences will be studied to distinguish their sentiment polarity [15]. Finally, the objects of sentence that expressed an opinion might be extracted. Opinion mining typically works on only positive and negative sentiment instead of working on discrete feelings and emotions such as happiness and sadness.

Discrete emotions would not distinguish sentiment strength but they could help to enhance the accuracy of association of words with positive or negative sentiments. Most of the opinion mining algorithms use machine learning techniques to classify general features related with positive and negative sentiment where the features could be a subsection of the words in the document, part of speech and so on. Two machine learning challenges for sentiment are feature selections and classification algorithm choices. Feature selection is defined as processing data to eliminate the least useful n-grams in order to improve the accuracy of classifications, which will be described further. Sentiment analysis also could be described as a method for Natural Language Processing. NLP or NLU, Natural Language Understanding, is the subset of computational linguistics, which itself is a combination of linguistics and computer science. In study of sentiment analysis, familiarity with some linguistic term can greatly help. Here, a few of such terms based on have been defined.

In [6] there are several research topics that are closely related to our research. These topics are online health communities. The objective in rundown is to extricate valuable data from the vast information accessible in the sites and different websites. Vinod L. Mane working for age of valuable data from the immense measure of information accessible in the distinctive sites likepatientslikeme.com, healthboards.com, and so on and condensing them for the following stage. AlokRanjan Pal streamlined lesk calculation which was one of the synopsis procedure utilized. Nandita Rane worked for the affiliation manage digging for type 2 diabetes. Calculation for mining affiliation rules have been utilized for removing data from the posts [12]. Utilized Sentiment examination which consists of in building a framework to collect and look at feelings regarding the item made in blog entries, remarks, surveys or tweets. Feeling examination focuses on states of mind, though customary content mining centers around the investigation of certainties. Content synopsis in light of scoring procedure where the word, sentence and diagram based scoring are assembled to add on some weight. ROUGE is the assessment technique which is utilized for checking and choosing number of sentences.

In [8] A substantial variety of cross-industry Twitter-focused publications exist, many exploring opportunities to use Twitter for forecasting financial and stock market outcomes, politics and election trends, civil unrest outbreaks, education and as a teaching aid in the classroom, sports, and another group of these directly related to approaches to text mining and sentiment identification. Far fewer related directly to understanding the potential of Twitter for healthcare-related applications such as disease surveillance, information dissemination, interaction between providers and patients, or gathering patient-level demographic and lifestyle behaviors are represented [13]. Although it might seem obvious that if Twitter data is useful in the variety of content areas listed above, the point of research is to test whether or not the assumption is valid.

In [10] Past examinations thought of specialized answers for separate client assessment on government health checking framework [8], influenza [9]. Despite of the extensive literature, none of the solutions have how forum relationships affect network dynamics. In the proposed framework, supposition investigation is finished utilizing Natural Language Processing, which characterizes a connection between client posted tweets and conclusion on the medication, and in addition, suggestions of much better medicines can be provided to the users. Accordingly the disadvantage of recognizing a connection between client post and the feeling related with the post is survived.

In [11] Surveys and behavioral risk factor surveillance systems such as those utilized by the CDC are commonly used methods to collect health data. However, it can take weeks or years to collect, clean, and analyze the data (Wartell, 2015). Given the dynamic nature of Twitter, it offers researchers a new method to collect relatively real-time data (Eichstaedt et al., 2015). By using data from Twitter, we are able to identify sentiments and topics using sentiment analysis and topic modeling [14]. The automatically identified topics and corresponding opinions provide a fine-grained understanding of opinionated text data that is done through social media to identify public opinion.

III. Background Study

Twitter is an online administration which gives endorsers a chance to post short messages ("tweets") of up to 140 characters about anything, from good morning messages to political stands. Such smaller scale writings are a valuable dig for getting a handle on feelings of gatherings of individuals, conceivably about a particular theme or item. This is much more in this way, since tweets are related to a few sorts of meta-information, for example, topographical directions of where the tweet was sent from, the id of the sender, the time data that can be joined with content examination to yield a significantly more exact picture of who says what, and where, and when. The most recent years have seen a tremendous increment in inquire about on creating sentiment mining frameworks of different sorts applying Natural Language Processing strategies.

One of major heart diseases with the high morbidity and mortality in India is Myocardial Infarction. Most MI causes by lack of blood brings to the heart, it leads to obstruction of coronary artery, chronic ischemia and hypoxia, the hemo-dynamic deterioration may leads to sudden death. MI diseases can only observed in some specific frequency resolution/scales, on other hand we gather the information through twitter tweets, that gains the interest of different treatments given or different types of medications given along specific period. To conquer these limitations, we propose a new MI detection and tweet classification approach based on the concept of deep learning. Rather than directly optimizing the sentiment analysis.

Cardiovascular illness is a worldwide general medical issue adding to 30% of worldwide mortality and 10% of the worldwide infection burden.1,2 In 2005, from a sum of 58 million deaths around the world, 17 million were the reason of cardiovascular ailment and, among them, 7.6 million were because of coronary heart disease.3,4 Myocardial dead tissue (MI) is one of the five principle indications of coronary illness, in particular stable angina pectoris, shaky angina pectoris, MI, heart disappointment and sudden demise. The expression 'intense coronary disorders' incorporates non-ST-elevation MI, precarious angina, sudden cardiovascular and ST-elevation MI passing. In epidemiological examinations, the frequency of MI in a populace can be utilized as an intermediary for evaluating the coronary illness load.

The new significance of MI by the World Health Organization (WHO) ought to encourage epidemiological checking, coding of the clinical analysis, legitimacy of death declarations and disease classification. Such an institutionalized case meaning of MI is of unique significance since it is a way to get dependable and practically identical information for assessment of the adequacy of anticipation and remedial techniques in nations with broadly fluctuating wellbeing frameworks. The definition has suggestions for the study of disease transmission, infection observing, substance of registries, clinical research ponders, clinical preliminaries, quality confirmation, monetary examination, medico-legitimate debate and estimation of social insurance costs. At the individual level, the finding of MI majorly affects physical and mental wellbeing and regularly on family, legitimate and protection matters.

IV. Proposed Methodology

A. Sentiment Analysis on MI

We have developed a method to gauge the Measure of Concern expressed by Twitter users for public health specialists and government decision makers. More specifically, we developed a two-step sentiment classification approach. Firstly, personal tweets are distinguished from News tweets. News tweets are considered as Non-Personal, as opposed to Personal tweets posted by individual Twitter users. In the second stage, the Personal tweets are further classified into Personal Negative tweets or Personal Non-Negative tweets. The two-step sentiment classification problem addressed in this chapter is different from the traditional Twitter sentiment classification problem, which classified tweets into positive/negative or positive/neutral/negative tweets without distinguishing Personal from Non-Personal tweets first. Although News tweets may also express concerns about a certain disease, they tend not to reflect the direct emotional impact of that MI disease on people.

The aim of sentiment analysis (or opinion mining) is detecting someone's attitude, whether positive, neutral, or negative, on the basis of some utterance or text s/he has produced. While a first step would be determining whether a statement is objective or subjective, and then only in the latter case identify its polarity, it is often the case that only the second task is performed, thereby also collapsing objective statements and a neutral attitude.

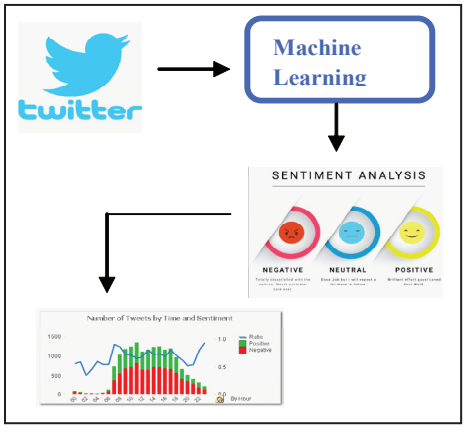

Fig. 1: Application Architecture

The language of our tweet messages is English, and thus the text processing is a easy process. As many tools that are freely available for English we can easily finished the analysis through it, we had to implement them in order to improve our model's performance. Our system uses logistic regression for classification, where words are weighted using TF.IDF (Term Frequency * Inverse document frequency).

B. TF.IDF

In information retrieval, term frequency – inverse document frequency or TFIDF, is a statistical intended to reflect the word is how important to a document in a collection of dataset. Weighting factor searches of information retrieval with text mining and user remodeling purposes. Incase,

In information retrieval, tfidf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. Suppose that we have term count tables of a corpus consisting of only two documents, as listed on the right. For pre-processing, we applied various algorithms, which we combined in order to achieve better performance. The pre-processing of the tweets goes as follows:

Table 1: Word Appearance of Document 1

Document 1	
Term	**Term Count**
Many	1
It	1
An	2
Example	1

Table 2: Word Appearance of Document 2

Document 2	
Term	**Term Count**
That	1
Was	1
Any	2
Some more	3

The calculation of tf.idf for the term "that" is performed as follows:

In its raw frequency form, term frequency is just the frequency of the "that" for each document. In each document, the word "that" appears once; but as the document 2 has more words, its relative frequency is smaller.

tf ("that",d1 = 1/5 = 0.2

tf("many",d2 = 1/7 = 0.14

An idf is constant per corpus, and accounts for the ratio of documents that include the word "that". In this case, we have a corpus of two documents and all of them include the word "that".

Idf("that",D)= log (2/2) = 0

So tf–idf is zero for the word "that", which implies that the word is not very informative as it appears in all documents.

Tfid("that",d1) = 0.2 * 0 = 0

Tfid("many",d2) = 0.14 x 0 = 0

A slightly more interesting example arises from the word "example", which occurs three times but only in the second document:

Tf("somemore",d1) = 0/5 = 0

Tf("somemore,d2)=3/7 = 0.429

Idf("somemore",D)=log(2/1)=0.301

The twitter users are insignificantly male, and biased with the respect to race distribution and civilization distribution, it depends on the fraction of all tweets that is classified as positive and negative tweets, from the above term count mechanism we found the tweet about MI, that makes an representative one, throughout this we find the mechanism of TF.IDF gives an best result than existing mechanism.

VI. Conclusion

Twitter has been as of late used to foresee or potentially screen true results, and this is additionally evident f or well being related subject. In this work, we extract information about diseases from Twitter with spatio-temp oral constraints, i.e. considering a specific geographic area during a given period. This is a very challenging task and needs to take care of a number of factors. Mostly the work related to the prediction and figuring out the heart problem, many data driven techniques has been used in past and the work inclines towards the classification problem. This is a process used to tune a model and then predict the class for whether the patient is suffering from any heart related problem or not.

The accuracy of the machine learning methods that we select to help the doctor take a decision rather, we want to decrease and penalize the model for having a bad prediction for the cases where the patient has a high probability for the heart attack but the model predicting for no heart problem.

References

[1] Neethu M,S, Rajashree R,"Sentiment Analysis in Twitter using Machine Learning Techniques", 4th ICCCNT 2013,at Tiruchengode, India. IEEE – 31661.

[2] Go, R. Bhayani, L.Huang,"Twitter Sentiment Classification Using Distant Supervision", Stanford University, Technical Paper, 2009.

[3] V. M. K. Peddinti, P. Chintalapoodi,"Domain adaptation in sentiment analysis of twitter," In Analyzing Microtext Workshop, AAAI, 2011.

[4] Long-Sheng Chen, Cheng-Hsiang Liu, Hui-Ju Chiu,"A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) pp. 313–322.

[5] Kennedy, D. Inkpen,"Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, Vol. 22, pp. 110–125, 2006.

[6] Qingliang Miao, Qiudan Li, Ruwei Dai,"AMAZING: A sentiment mining and retrieval system", Expert Systems with Applications 36 (2009) pp. 7192–7198.

[7] Pang B., Lee L.,"Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," In Proceeding of the 43rd Annual Meeting of the ACL, Stroudsburg, 2004, pp. 221 of 619.

[8] Tang H., Tan S., Cheng X.,"A survey on sentiment detection of reviews," In Expert Systems with Applications: An International Journal, Vol. 36, pp. 10760-10773, 2009.

[9] Shailesh Kumar Yadav,"Sentiment Analysis and Classification: A Survey", International Journal of Advance Research in Computer Science and Management Studies Vol. 3, Issue 3, 2015.

[10] Kennedy, D. Inkpen,"Sentiment classification of movie reviews using contextual valence shifters, Computational Intelligence", Vol. 22, pp. 110–125, 2006.
[11] Pang, B., Lee, L.,"Opinion mining and sentiment analysis", Foundations and Trends® in Information Retrieval, 2(1–2), 1-135, 2008.
[12] Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.,"Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques", In Proceedings of the Third IEEE international conference on data mining, 2003.
[13] A. Go, R. Bhayani, L. Huang,"Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, 1, 12, 2009.
[14] A. Pak, P. Paroubek,"Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In LREc, Vol. 10, pp. 1320-1326, 2010.
[15] M. Bilal, H. Israr, M. Shahid, A. Khan,"Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques", Journal of King Saud University-Computer and Information Sciences, Vol. 28, No. 3, pp. 330-344, 2016.
[16] W. Medhat, A. Hassan, H. Korashy,"Sentiment analysis algorithms and applications: A survey", In Shams Engineering Journal, Vol. 5, No. 4, pp. 1093-1113, 2014.
[17] Alexander Pak, and Patrick Paroubek,"Twitter as a Corpus for Sentiment Analysis and Opinion Mining", In LREC-2010, Valletta, Malta.