Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Enhanced Exploration of Sentimental Analysis in Health Care

Kannan Chakrapani
SASTRA Deemed University
Muniyegowda Kempanna
Bangalore Institute of Technology
Safa Mohamad Iqubal
SRM Institute of Science and Technology
Kavitha Thyagarajan
Koneru Lakshmaiah Education Foundation
Manikandan Ramachandran
SASTRA Deemed University: Shanmugha Arts Science Technology and Research Academy
Vidhyacharan Bhaskar (meetvidhyacharan@yahoo.com)
San Francisco State University https://orcid.org/0000-0003-3820-2081
Ambeshwar Kumar
SASTRA Deemed University: Shanmugha Arts Science Technology and Research Academy

Research Article

Keywords: Sentimental analysis, Sanitization, Tokenization, Stemming, lemmatization, LDA, Machine learning classifiers

Posted Date: July 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-620229/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

AN ENHANCED EXPLORATION OF SENTIMENTAL ANALYSIS IN HEALTH CARE

Kannan Chakrapani
School of Computing
SASTRA Deemed University
Thanjavur 613401
Email: kcp@core.sastra.edu

Muniyegowda Kempanna
Computer science &Engineering Department
BIT, Bangalore
Email: kempannam@bit-bangalore.edu.in

Mohamed Iqubal Safa
Assistant Professor
Department of IT, SRM Institute of Science and Technology
Kattankulathur Chennai.
E-mail: safam@srmist.edu.in

Thiyagarajan. Kavitha
Department of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation,
Greenfields, Vaddeswaram, Guntur-522502
Email: drtkavitha@kluniversity.in

Manikandan Ramachandran
School of Computing,
SASTRA Deemed University,
Thanjavur, Tamil-Nadu, India
Email:srmanimt75@gmail.com

Vidhyacharan Bhaskar
Dept. of Electrical and Computer Engineering,
San Francisco State University,
San Francisco, CA 94132, USA.
Email:vcharan@gmail.com

Ambeshwar Kumar
School of Computing,
SASTRA Deemed University,
Thanjavur, Tamil-Nadu, India
Email: ambeshwar.kumar@gmail.com

Abstract

The medical dataset replicates the patient's crucial information, such as important details regarding the patient's health. It includes disease diagnoses, interventions, and descriptions of the examined results. Also, detecting the mindset of an acute disease-affected patient is a primary challenging task. Though sentiment analysis plays a role in seeing their perspective, the significant broad medical application does not yet meet the analysis of patient mindset. So here we identified major shortcoming exists while studies the diversified disease-affected people mindset. Hence, we introduce a practical framework to analyse patients' perspectives using a socio-medical dataset that contains various reviews and feedback of critical diseases-affected people—initially, we applied a pre-processing technique, including Lowercase Conversion, removing special characters, removing stop words, Number to word conversion, Stemming, and lemmatization over dataset. Next, N-gram tokenization methodology is used to extract the valuable features followed by assigning polarity score to each sentiment we extract and calculate the overall polarity of the context. Finally, a probabilistic LDA model was employed to combine the review. Furthermore, various machine learning classifiers are explored to evaluate the performance of the proposed framework.

Keywords: Sentimental analysis, Sanitization, Tokenization, Stemming, lemmatization, LDA, Machine learning classifiers.

1. Introduction

The evolving area in medical concepts depends on allotting the classification and emotion analysis. Because of the absence of domain-specific lexicons and uninterested domain researchers in this area, the issues accepted are high. One more issue is the semantic relation of the healthcare domain and the separation of knowledge-dependent features. The main reason is that the all-time medical lexicons do not provide any characteristics like category and sentiment. The experts are tried their best to plan the data extraction like GENIAI and PennBioIE 2 to solve the issues throughout different the years. . The primary need is to recreate either structure or unstructured corpora versions. Other economic and ontology methods are used with linguistic and ML (machine learning) techniques [1]. These are used to separate the healthcare topics with the synaptic and

semantic characters [2]. Two systems are evolved to separate the semantic relations from the healthcare domain in recent work. The first system is tokenization (allotting the groups). The second system is for recognizing the emotions in the healthcare domain and their contexts. Healthcare topic is nothing but a phrase or a word with entities and knowledge and data belonging to healthcare attributes. The recognized field consists of two types i) Medical ii) Non-Medical. Separations of negation and stop words or sentences are considered with the aim of identifying the area. Let us take two examples Regular headache and uncontrolled jerking. These are considered medical and non-medical fields depend on the presence of the healthcare domain. Headache is regarded as a symptom of starting stage of cancer also. So it is a medical context with a company of medical concepts. Additionally, every word or phrase of the corpus is recognized as the context in our work. In case of sentences like" Orange is good or bad" means a non-medical domain without any medical concepts

The categorization and emotion recognition systems are used to separate the contexts and domain. The separated field is divided into five models in the categorization. I) Disease II) symptoms III) drugs IV) human anatomy V) miscellaneous medical terms (unidentified topic represented as MMT in the remaining section of the paper). An example of a disease type is "Headache." According to utterance and computation of the first incident of separation concept, the healthcare researchers and scientists presented these five categories in the corpora. Every type has its ideas which allot the whole classification based on their context. Eleven classifications of healthcare domain are recognized in pairwise aspects like disease symptom, disease drug, etc. According to both concepts and contexts, the sentiment recognition model [3, 4, and 5] is augmented for sense-based information. Here, only the positive and negative emotions are taken into account. For example, "There is something wonderful about being pregnant" is represented with positive emotion. The outputs of emotion recognition are not similar for various types of domain. Consider "anatomy of the human body" as a neutral sentiment while positive or negative feelings depend on symptom types. The past model lexicon viz is used to evaluate these models. WordNet of Medical Event is utilized to separate the healthcare domain from contexts. Additionally, the lexicons allotting the linguistic and the emotional characteristics in the healthcare domain [6], the WME has two versions: WME 1.0(WME version 1) [7] and WME 2.0 (WME version 2) [8] augmented to compromise. The Linguistic character such as POS, gloss with polarity

score, emotions are coming under the WME1.0 lexicon. It also has 6415 no. of healthcare domain, and it impotent to offer the feeling-based field and related knowledge data. So we go for WME 2.0, which has 10186 no. of the area, and it has some knowledge-related character like affinity score, similar sentiment words (SSW), and gravity score. The blended model is injected to combine the previous linguistic symbols of WME 2.0 and the machine learning prototype. This offers aspects such as negation [9], uni-gram and bi-gram. It is two types of classifiers to accomplish an average of 0.81 and 0.86 F-Measures for allotting classifications to the healthcare domain and contexts in developing a categorization system. i) Naive Bayes ii) Logistic Regression [13], In the presence of WME2.0, the emotion recognition system was evaluated with the utilization of naive Bayes and support vector-oriented Sequential Minimum Optimization (SMO) Classifiers. It attains an average of 0.91 and 0.81 F-Measure for recognizing emotions of the healthcare domain and contexts. When the unigram and bigram are used to identify the divisions of the medical domain, the negation character identifies the first sentiments of medical concepts [10, 11, and 12].

The paper is organized as follows: **Section 2** discussed the various related work and literature studies associated with sentimental analysis. **Section 3** discussed the current problem statement along with the solution. The different methodology used for the proposed approach is discussed in **Section 4**. Experimentation and result discussion was on **Section 5**. Finally, the paper is concluded in **Section 6.**

2. Related Study and Literature Survey

This section analyzes and describes the numerous existing approaches built for analyzing sentiment of acute disease affected patients.

Many new researchers are examining the way social media convincing the public of medical care. Researches are injecting text mining (Ficek and Kencl[14], Rahnama[15]), and it is doing a significant role in the high performance of unstructured data by the utilization of apache spark along with the binary and ternary process, Baltas and Tsaklidis[16] introduced a Twitter sentiment analysis. The following method is a conventional extreme learning machine based on spark cluster performed by Oneto et al. [17]. Using spark with deep learning shows that they have a high-performance level than any other spark model. In mobile big data analytics, a deep learning framework is introduced by Chen et al. [18] utilizing the apache spark model. The further method

is sentimental analysis with spark architecture on large scale data discovered by Nodarkis et al. [19].

To separate the sentiments from the HPV vaccine-based tweets, the best ML system is performed by Du.et.al [20]. The ranking division, along with the SVM type, is processed, and 6000 tweets are explicated physically. Other than baseline types, this output gives the best of 0.6732F-. Medical sentiment analysis is an evolving technology. Denecke and Nejdl[21] introduced a health care ontology to evaluate the factual level in the healthcare texts. It is different from past emotion examination systems. Emotion examination is processed by rule or by ML methods. Many workers are represented with ML methods than rule-dependent methods.

To find out the polarity level in sufferer data, Xia et al. [22] discovered a multistep opinion classification. Additionally, to compute the quality of healthcare, Cambria et al. [23] analyzed an outline by combing sentic PROMs and sentiment examination systems. Disease like breast cancer, colorectal cancer, and diabetes are classified by De la Torre-Diez et al. [24]. Social media depends on online cancer group relations; portier et al. [25] imposed emotion examination methods to find false sentiments and bad moods in a human. The subsequent research examines the emotions of a group of people attempted by Crannnell et al. [26]. Sufferers mentally convince these people. Extraction of online e-liquid revises by e-liquid characters by Chen and Zeng [27]. It is to separate the polarity characters and also process an emotion examination in big online e-liquid websites.

Ozcift and Gulten [28] introduce the research of ML algorithms in healthcare identification. To evaluate the division execution, they integrate the one ML classifier, including the CFS algorithm. 74.5%, 81%, and 87.2% percentage of output offered by three healthcare datasets. It is compared to basic classifiers. The following model is CNN-MDRP by Chen et al. [29][3] to forecast the infection complication from unstructured and structured information. This model gives a 94.8% percentage of convergence speed when it happens in actual life incidents. Depend on data categorization; Lu [33] introduced a concept recognition type. They perform character depend on categorization by gathering information from internet medical groups. It uses various division models along with C4.5, SVM, and Naive Bayes. When compared to other methods, it gives the best-improved classification outputs.

A distinctive model for refining word-level emotion examination by Chen et al. [29,30] and gives the best emotion analysis results than the old eleven methods. Lin et al. [31] introduced TCM medical documents. They got multi-combination type by integrating various characters sing weighted LDA concept type. The effectiveness of this work was evolved along with the best categories rate. It recreates the best support in TCM medical science. Jonnalagadda et al. [32] transverse research on recognizing decision experts from 178,527 news research papers, the output gives 88.5%of efficiency for the 734,024 samples along with corpus. Monogram et al. [34] introduced the best deep learning model for cardiovascular disease with many kernel training. The neural language like CBOW and Skip-gram is imposed by Minarro-Gimenez et.al. For the first time using the skip-grams, it also uses PubMed corpus and PubMed text articles. Th et al. [36] performed skip-gram and CBOW on 1.25M PubMed research papers to evaluate the word combination with the twinning couple. For the purpose of biomedical NLP, Chiu et al. [37] processed best word embedding's and performed two various corpora. It demonstrates skip-gram model surpasses the CBOW model. From the suffering with anorexia nervosa, Spinczyk et al. [38] offered a rule-dependent type for examining emotions. The emotional word is recognized from the record. At the time of healing, it assists humans to the sharp using group of phrases model.

3. Problem Statement and Solution

From the literature survey, we found the difficulties involved while analyzing the acute disease-affected mindset. During the extraction of sentiment, there have been a lot of parameters involved. Also, analyzing the diversified people mindset becomes a very challenging role. Hence, we proposed a practical framework to analyze the sentiment of perceptive disease-affected people. In our proposed framework, we used five majorly used machine learning classifiers known as Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), K-nearest neighbor (KNN). Our proposed framework includes:

3.1. Initially, we perform pre-processing over the dataset, including Conversion of Lowercase, Eliminate special characters, Eliminate stop words, Conversion of Number to word, Stemming, and lemmatization. Hence the dataset gets accurate and crisp.

3.2. After pre-processing, N-gram tokenization is performed over the dataset.

3.3. Then assigning a polarity score to each extracted review and calculate the overall polarity score.

3.4. Through combing the data review, we employed a probabilistic LDA over the resultant dataset.

3.5. Then primary machine learning classifiers such as Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), K-nearest neighbour (KNN) are used to classify the sentiment as positive, negative, or neutral.

4. Methodologies

This section will figure out all the methodologies used for our proposed work and describe its entire functionalities in a detailed manner.

4.1. Data pre-processing: Data pre-processing is primary task of any data classification process. It is very crucial as well as much generalized approach. Here we used four major pre-processing technique is employed over our socio-medical dataset. It is helpful to improve the quality of classification process as well as paves the way to develop the features as much robust.

a. Conversion of Lowercase: Initially, we employed lowercase conversion. It sequentially scans the entire dataset and finds out the uppercase words. If the uppercase word exists, it converts it into corresponding lowercase letters using the python library function known as Numpy.

b. Eliminate special characters: This is the second of data pre-processing. This step eliminates the special characters such as (*, %, $, @, #, etc.) from the dataset. And it takes the uppercase-free dataset, which means after finishing step 1, it starts its process.

c. Eliminate stop words: Usually, stop words are frequently occurring in the text, and no meaningful information is conveyed. Hence my use of the NLTK library function we achieved to eliminate the stop words. Fig. 1 demonstrates the clear view.

d. Conversion of Number to word: These steps involved converting the number to words by use of python library known as num2words. (For instance, 7 are converted into seven).

e. Stemming and lemmatization: It is the final step of pre-processing. Stemming is the technique to remove the suffix or prefix from a given the word and reduce the word's complexity, eliminating the stem or root of a term. We employed a porter stemmer to done this process. Lemmatization usually diminishes the word to a well-founded dictionary word.

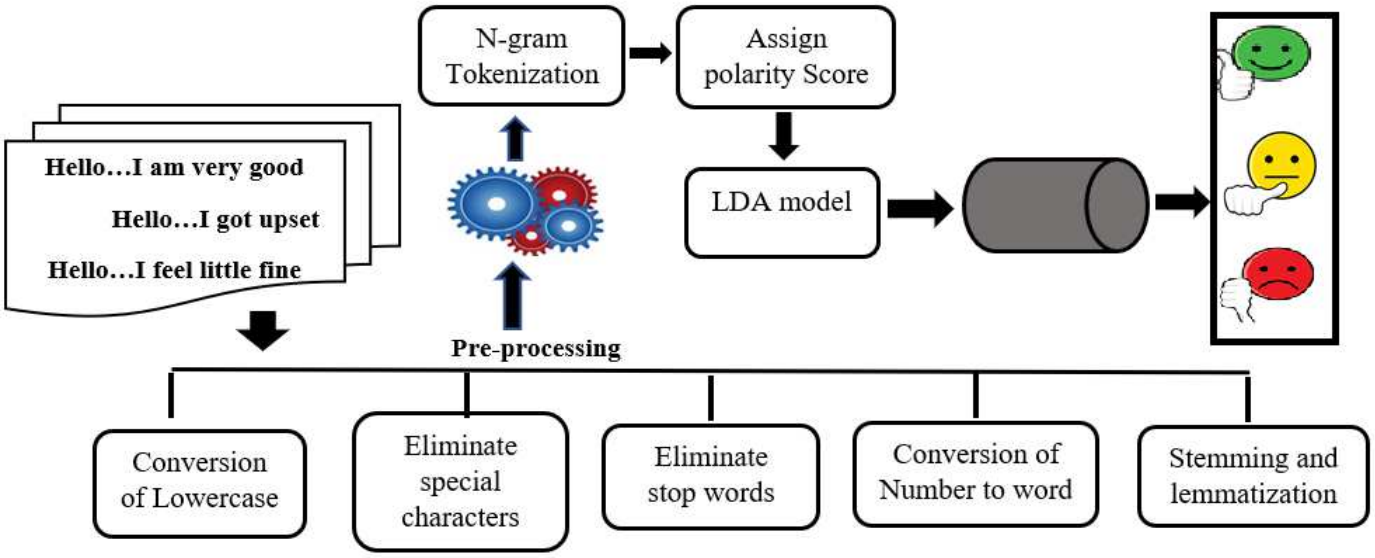


Fig. 1: Proposed workflow

4.2. N-gram Tokenization:

N-gram tokenization [39, 40, and 41] is defined as a collection of occurring tokens present in a frame and used to find further grants. It divides the text into different tokens after the pre-processing with the intention of feature extraction.

Fig. 2: Token analysis

Based on the pre-processed data, the N-gram dictionary is explicated physically at words like 'a',' and," the," there,' etc.; these words do not have any critical data .so they are rejected from the review sentences. Separation of the opinion words from the N-gram dictionary is processed by sentence-level annotation and summarization. The summary of every sentence, along with frequency, is represented in fig 2. Using the summarized characteristics, the mark of every aspect

was evaluated to predict the no. of positive, negative, and neutral words is represented in table 1. Also, it has to be done after the physical explication and summarization. The score was calculated for positive, negative, and their emotions individually. Additionally, it has three steps i) recognizing the opinion terms ii) feature vectorization process iii) vector transformation. Depending on Term Frequency and Inverse Document frequency [TD-IDF] (42), the features of vectors are computed. The TD-IDF allots the weight to every feature vector. The importance of the feature vector is calculated by

$$\text{Td-Idf}(x,y,Y)=\text{Td}(x,y).\text{Idf}(x,Y) \text{ ----(1)}$$

$$\text{Idf}(x, Y) = \text{Log} \frac{n}{|\{y \in Y : X \in Y\}|} \text{ ------- (2)}$$

Where,

X - Entire time a specific term takes place in a document ‘y’

N – Entire number of documents in corpus

$|\{y \in Y : x \in y\}|$-no of time the term ‘x’ happens in total number of documents ‘Y’

The results of the weighted character vector are fashioned to LDA (latent Dirichlet allocation) [43] using pipeline. Intentionally, the Bayesian optimizer which is in LDA separates the some characters by changing them into various no.of concepts. ‘Dimension is term used to describe every concept in LDA. Below section briefly describes the LDA working type.

4.3.Assign Polarity Score:

Depends on the polarity level of total domain in context, emotions are bring out by learning section[44].some phrases like “no” ,”not”, ”never” and” neither” are considered to identify the relevant emotions of healthcare domain[45-47].The below algorithm is sketched to allot the emotions to healthcare domain by utilizing the emotion recognition system.

STEP 1: Set polarity score (Polarity $_{level}$) and emotion of both healthcare and non-healthcare of the domain. The different emotion lexicons techniques used for imposing to allot the polarity level are SenticNet and SentiWord Net.

STEP 2: In order to allot the relevant emotion of domain, determine the negation words or phrases

STEP 3: Calculate the total polarity level of the domain using following equation,

$$\text{Polarity}_{level} = \sum_{n\equiv 1}^{k} Polarity_{level} \text{ ---------- (4)}$$

Where,

Polarity $_{level}$ - the total polarity level of the context

$Polarity_{level}$ - single polarity level of every topic in a domain

The following algorithm is to examine the single person's emotions. It should be done after the polarity level determination. It can find the classification of healthcare domain along with domain classification.

STEP 1: Set the type of healthcare domains of the environment with concept categorization system. We can denote the medical concepts and their types as CM in an environment.

STEP 2: A new abbreviation Pcc is introduced which is used to denote the successive healthcare domain and their types

a. **If** the successive twin partner of concept types are same then Pcc is

$$Pcc = CM1 \cap CM2 \text{ ----- (5)}$$

b. Else

$$Pcc = CM1 \cup CM2 \text{ ------- (6)}$$

Where CM1 &CM2 – two successive healthcare domain and their types in a environment

c. In order to find the total context category(Cc), use extracted Partial Context category(pcc)

$$Cc = Pcc1 \cap Pcc2 \text{ --------- (7)}$$

Where Pcc1 &Pcc2 – partial context category of the environment

4.4. Latent Dirichlet Allocation (LDA) with Topic modelling:

Following hierarchical Bayesian type, LDA is injected on character vectors to inspect the text in the corpus. The unsupervised likelihood type that designs corpus into the group of content is called LDA [48]. Over the words, each concept is sketched as a distribution. Consider there is possible arrangement 't' along with the corpus 'c', and it holds 'r' number of review associated with it. Each probability distribution function associated with the feature vector is considered the polynomial probability distributed function, and each study will produce the arbitrary constant 'k'. The following equation represents how feature extraction is done in the LDA model.

$$p(f_a) = \sum_{b=1}^{n} P\left(\frac{f_a}{t_a} = b\right) P(t_a = b) - - - - - (8)$$

Where,

P (t_a= b) -the likelihood concept of b sampled for character fa for every revise in corpus c.

P ($f_i|t_i$ = k) - the likelihood of fa under topic b and n denotes the total number of domain.

The term had used while computing equation 8. There we used a character vector that refers to b, which is also a multinomial distribution of characters. For considering the review r, P (t) refers to the multinomial distribution. For the representation of feature vector and reviews the estimated parameters, and are used, the total features representation of a review is done using NR. Also, R represents the entire set of reviews. The topic-word hyperparameters are represented using, and the total consideration associated with Dirichlet distributions is kept on updating as every unit of the cell.

θ Determines the variable comprises of review level and it is sampled once per feature level variable i.e. R and f corresponding to each work r with N. Consequently, for entailed features also very difficult to score and process directly, equation (8) clearly determines the probability function of overall possibilities available with respect to each feature vector.

$$P(t_i = k|t_{-i}, f_i, r_i, \ldots\ldots)\alpha \frac{C_{fi}^{FN} + \beta}{\sum_{f\equiv 1}^{F} C_{fj}^{\mathrm{FN}} + F.\beta} \cdot \frac{C_{rij}^{RN} + \alpha}{\sum_{k\equiv 1}^{N} C_{rj}^{\mathrm{RN}} + N.\alpha} \text{ ----- (9)}$$

Where,

$t_i = k$ - The f_i features which is assigned to the topic k.

t_{i-1} – The i^{th} review represents to the allocated domain $f_i * r_i$.

R – Entire set of reviews.

F – Entire set of features.

C^{FN} and C^{RN} - Matrix corresponding to topic-review.

C_{fj}^{FN} – The overall features corresponding to topic k.

$f_i * C_{rj}^{RN}$ – The respective topic given to some word with respect to review r without f_i .

The following equation (8) and (9) represent the parameter θ and $\hat{\phi}_i^{(r)}$

$$\hat{\phi}_i^{(k)} = \frac{C_{fi}^{FN} + \beta}{\sum_{f\equiv 1}^{F} C_{fj}^{\mathrm{FN}} + F.\beta} \quad ---(10)$$

$$\hat{\phi}_i^{(r)} = \frac{C_{rij}^{RN} + \alpha}{\sum_{k\equiv 1}^{N} C_{rj}^{\mathrm{RN}} + N.\alpha} -----(11)$$

In this proposed work, LDA is utilized to discover the text from the review corpus and fuse them into the latent domain. We have modeled the field as K = 100, 200, 300, 400, and 500 on the review corpus. The essential features were identified from the domain based on the probabilities correlated with each segment. Later, the implementation of feature selection is presented in the below section to handle the curse of dimensionality problem.

4.5.Classification process:

In this subsection, we discussed various classifiers used for our evaluation purpose; by using those classifiers, we analyze and classify the sentiment as positive, negative, and neutral.

a. Support Vector Machine (SVM):

While text classification and important categorization are based on hypertext, the SVM model is widely used. It is very much helpful to significantly minimize the training sample label of both inductive and transductive settings. It is a classifier that primarily classifies the cost function also enhances the classification performance. And we have used the LibSVM library, which is a function of sklearn SVC. To analyse the data in minimizing the structural risk, a learning robust method, SVM, is used. It is a kind of learning method in which classifies the data appropriately.

The training phase was usually optimally separating hyperplane, reducing the cost function so that distance between two classes of margin had been induced and feature space must be minimized. Consider there is m instance of data in the training phase. Each model consists of a pair (ai, bi) where ai∈ Rn is a vector attribute that belongs to the instance i. Also, bi ∈ {+1,-1} and known as the instance of class label.

To find the hyperplane that finely separates the optimal solution, the main objective of the SVM and the corresponding data belongs to two main classes, which is W. a+y=0. The decision function is used to classify and test the instance y, which is defined as

$$F(y) = W.a+c \text{ ---------- } (12)$$

b. Naïve Bayes:

The Nb classifier between the predictors depending on Bayes theorem detachment, since NB refers to different characters, the multinomial naive Bayes are used along with proper fit prior. It is simple, has no difficulties to sketch, no repeated variable computation utilizing big datasets. It processes enlightened division models. Bayes theorem provides a way of calculating the posterior probability, $P(g|h)$, from $P(g)$, $P(h)$, and $P(h|g)$.

Naive Bayes classifier assumes that the effect of the value of a predictor (h) on a given class (g) is independent of the importance of other predictors. This assumption is called class conditional independence.

NB offers a computation of posterior probability p(g|h) from p(g),p(h) and p(h|g). It consider the performance of total of a predictor (h) on a given class (g). It is not dependent of other predictors. It is called conditional independence

$$P(g|h)=p(h1|g)*p(h2|g)*\ldots.. p(hn|g)*p(g) \text{ ------ } (13)$$

Where,

P(g|h) is the posterior probability of class (destiny) given predictor (attribute).

P(g) is the prior probability of class.

P(g|h) is the likelihood which is the probability of the predictor given class.

P(h) is the prior probability of the predictor.

c. Decision Tree (DT):

It comes from the classification tree algorithm family. Transversely, the subtrees are outlined by dividing the source entity. 12 is the maximum depth of the tree. The tree structure is the basic form for classification and regression models. It can be built by splitting the dataset into tiny and very tiny subsets. Simultaneously, the association tree is evolved. The decision node and Leaf node are the final results. As usual, the topmost node is the root node. The root node handles the categorical and numerical data. It is a top-down approach, so the data are divided into different homogeneous values which have instances. The entropy of the decision tree is built by the following equation.

$$E(S) = \sum_{j=1}^{d} -Pi\ log2\ pi \text{ ------ (14)}$$

d. Random Forest (RF):

Based on the dataset sample, the RF has various decision trees. There is a possible 'n' maximum depth of the tree. It can be built with different single decision trees at the learning phase. To make the destiny prediction, it makes use of predictions from all the trees. The mode of the classes for classification or the mean forecast for regression, since there are a group of outputs to reach a destiny, they are called an Ensemble method.

The following equation used for making binary tree:

$$Nlm=WmCl - Wleft(m)Cleft(l)-Wright(m)Cright(m)\text{-------(15)}$$

- sub(m)= the importance of node m
- w sub(m) = weighted number of samples reaching node m
- C sub(m)= the impurity value of node m
- left(m) = child node from left split on node m
- right(m) = child node from right split on node m

The importance for each feature on a decision tree is then calculated as:

$$F_i = \frac{\sum j{:}node\ j\ splits\ on\ feature\ i\ nij}{\sum K \in all\ nodes\ nik} \text{------ (16)}$$

Where j = the importance of feature m

K = the importance of node m

e. K- nearest neighbor:

K nearest neighbor is the faster algorithm which significantly produces the classification result as accurate and more precise also the performance must be enhanced. It is majorly used to find out similar objects and much applicable to solving the problem. Applications such as recommendation systems, search engines are utilized and essentially work based on KNN.

5. Experimentation and result Discussion

This section presents a clear description of the implementation details and a proper comparison with the existing approach. We used the Windows operating system with 12 GB RAM with a 2 GHz processor and 1 TB hard disk for experimentation purposes. Python programming language along with the help of proper library functions used for implementation. We used pycharm IDE for implementation purposes, and due to the lack of storage issues, the dataset needs GPU. Hence, we used google colab.

Dataset Description:

Notably, The 821,483,453 general tweets on Twitter are brought together between 16th march 2019 and 2nd October 2020. Among them, 438,072,932 are based on healthcare issues, especially numerous social environment health domains. Besides, three medical and health datasets are used to assess the coherence of the project. In the below sections, a brief explanation of this data is given. The other standard survey has taken between October 2013 to Jan 2016 using the UCI machine learning repository (common dataset in Twitter).This convention dataset has different medical tweets, which are gathered using many accounts on Twitter as follows: a. reutershealth b. kaiserhealthnews c. latimeshealth d. bbchealth e. msnhealthnews f. NBChealth g. cbchealth h. nytimeshealth i. gdnhealthcare j. everydayhealth k. nprhealth l. foxnewshealth. Table 1 describes the details regarding statistical data and Table 2 describes examples of various emo-tags list.

Aspect	Total count
Total emotag	7,45,21,43,542
Total Negative emotions	8,45,32,54,912

Total Neutral emotions	6,34,32.87,454
Stop words	8,34,23,12,343
Set of tags	10,23,43,34,123
Resources from URL	10,34,54,23,567
Number of total terms	15,34,34,23,232
Number of unique words	4,34,34,32,654

Table 1: Details regarding the statistical data

Positive	**Negative**	**Neutral**
:-)	:-)	:-O
:)	:\	:-J
8D	:'(	8-0
::P	:'-(	:-|
XD	D- ':	:|
:-D	:<	:-0
:3	:-c	:-||

Table 2: Example of emo-tags list

Here our experimentation consists of 5 main phases. It includes data pre-processing which provides for conversion of lowercase, elimination of special character, elimination stop words, and conversion of number to terms, and stemming and lemmatization. After the resultant pre-processing dataset is passed to N-gram tokenization that emotion has been converted into tokens. Then we assign polarity value to each emotion and calculate cumulatively. The resultant had fed to Latent Dirichlet Allocation (LDA) with Topic modelling. In that, each topic is converted into a set of sentences. Then finally significant classifier like SVM, NB tree, DT tree, Random forest, KNN is used to analyses the sentiment of acute disease-affected people.

Step 1: Data pre-processing

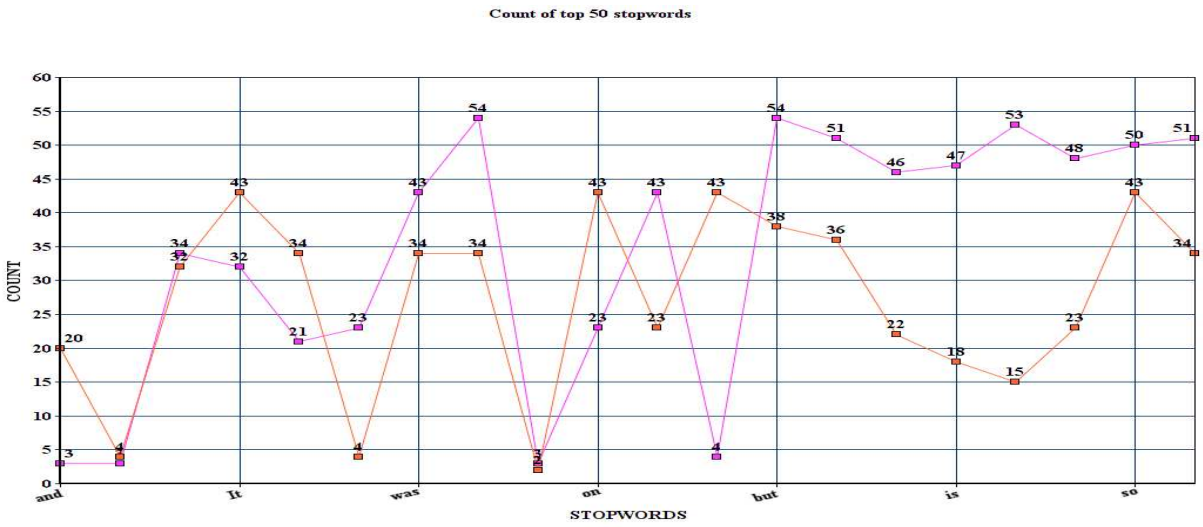
The collected dataset is applied over the set of pre-processing techniques. We also include some manually developed wordlist, which provides for some crucial keywords of sentiments like positive, negative, and neutral. Some of which are shown in the below table. We also embedded it in the pre-processing step.

Positive	**Negative**	**Neutral**
fitness	Menace	skin
patient	Not-safe	Cancer
likelihood	Viruses	tablet
Protection	Injury	syndrome
Natural	Misuse	failure
Prevention	Bacteria	Longevity
Doctor	Medicine	suffocating
Boost	Death	Dilemma
Hospital	Bat virus	choices

Table 3: Set of manually developed wordlist

Step 2: N-gram tokenization

Then we applied N-gram tokenization over the resultant dataset. There annotation based on sentence-level and summarization is done and extracting the words based on opinion from N-gram dictionary. The words such as “a”, “the”, “and”, “so” etc…kinds of words are removed from the dictionary and provide clarity about what kind of emotions we want to predict. The resultant of N-gram tokenization is simulated and generated the graph and shown in below. Pink line is our approach and orange link base line approach which is except N-gram tokenization [i.e. only pre-processing]. Hence, we found that our approach outperforms well and able to provide result in an accurate manner.



Graph 1: Generated after N-gram tokenization to dataset

Step 3: Assign polarity score and calculate it's cumulative

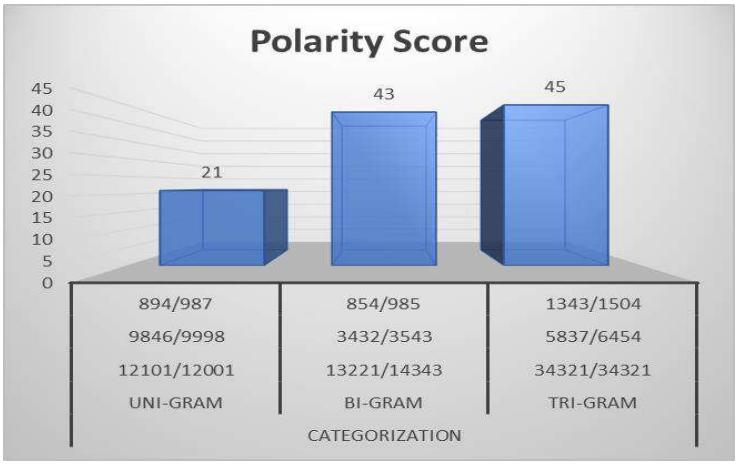
This step happens after the completion of N-gram tokenization. The resultant of n-gram tokenization is fed to the input of this step. To analyses the exact prediction, we will set some polarity scores to each entity so that the negation kinds of words such as "not," never," "none," neither," etc., can be recognized and extract the exact sentiment. After fix polarity score to every entity, we categorize this context into uni-gram bi-gram and tri-gram which defines the priority to analyses the sentiment also it is used to examine the kinds of statistics result involved in the disease.

Statistics	**Categorization**		
	Uni-gram	**Bi-gram**	**Tri-gram**
Disease	12101/12001	13221/14343	**34321/34321**

Symptom	9846/9998	3432/3543	5837/6454
Drug	894/987	854/985	1343/1504
Polarity Score	21	43	45

Table 4: Categorization of sentiments

The categorization over each entity is done exactly with Tri-gram in which the corresponding polarity score=45. After finding these, cumulative is done over each entity, then entire sequence must recognize the negation, which means the negative sentiment is done. The comparison graph over each categorization is shown in below graph.



Graph 2: Categorize sentiments based on polarity score

Step 4: Latent Dirichlet Allocation (LDA) with Topic modelling

It was assigning probability value over every emotion and categorizes that based on that probability value. Also, it categorizes that into a high risk to low risk. Furthermore, it keeps on actioned this, so that model fine-tuned most finely. Here k = 100.

Figure 3 demonstrates the overall entities of the dataset, which clearly scatters entire emotions involved, got from the LDA model. It fine-tuned and provides a high reliable feature in which classification done with high accuracy. Each dataset which are scattered in figure 3 is the emotions i.e. (Positive, Negative, and Neutral). Also table 5 and 6 describes the different emotions based on its random fixed probability values.

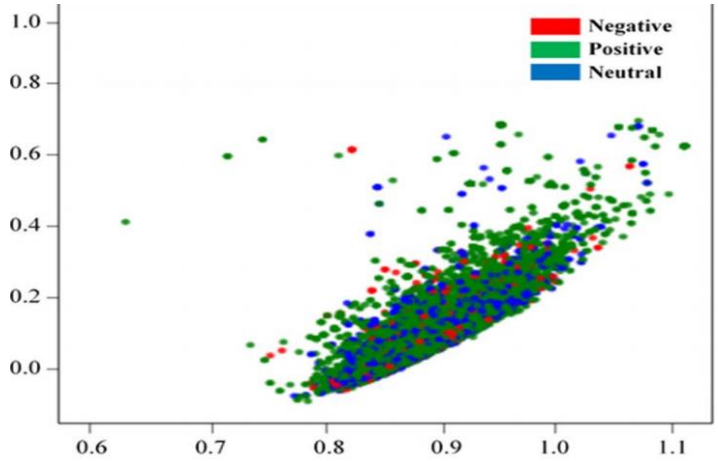


Figure 3: Overall entities of dataset

Emo-tags 100		**Emo-tags 300**		**Emo-tags 500**	
Emotions	**Probability**	**Emotions**	**Probability**	**Emotions**	**Probability**
Positive	0.003432	Positive	0.04342	Positive	0.08372
Negative	0.002342	Negative	0.00875	Negative	0.03721
Neutral	0.004543	Neutral	0.06563	Neutral	0.08374

Table 5: different emotions based on its random fixed probability values.

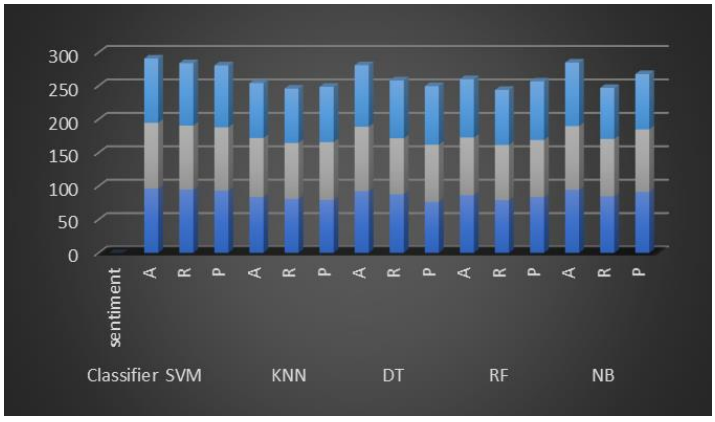
Step 5: Evaluate result using classifier:

For evaluation of result, we use most significant classifier like SVM, KNN, DT, RF, NB tree that are widely used for analyzing very crucial information. Classifier extracts the exact sentiment from each existing entity. We used evaluation metrics like Accuracy, Recall and Precision for our analysis.

Classifier / Sentiment	SVM			KNN			DT			RF			NB		
	A	**R**	**P**	**A**	**R**	**P**	**A**	**R**	**P**	**A**	**R**	**P**	**A**	**R**	**P**
Positive	**96.5**	94.9	93.4	83.6	80.9	79.4	92.5	87.9	76.4	86.4	78.9	83.4	94.5	84.9	91.4
Negative	**98.2**	95.6	94.5	88.3	83.6	86.5	96.2	83.6	85.5	86.3	82.6	85.5	95.2	85.6	93.5
Neutral	**96.1**	93.4	92.6	82.1	81.4	82.6	92.1	86.4	87.6	87.2	82.4	87.6	95.1	76.4	82.6

Table 6: Comparison of various classifiers for analyzing sentiments

From above table describes the overall comparison of the classifier to classify and analyses sentiment. Where A denotes Accuracy, R denotes Recall, P denotes precision. In our experimentation analysis, we determine, SVM Outperforms well and produce better accuracy when compared to the remaining classifiers. It achieves the highest accuracy of **98.2%** while classifying negative sentiment, **96.5%** for organizing positive feelings, and analysing neural emotions achieves **96.1%** accuracy. Next from SVM, Naïve Bayes outperforms well, and it also performs an accuracy of **95.2%** accuracy for classifying negative sentiment. Consequently, the corresponding comparison graph is given as follows.

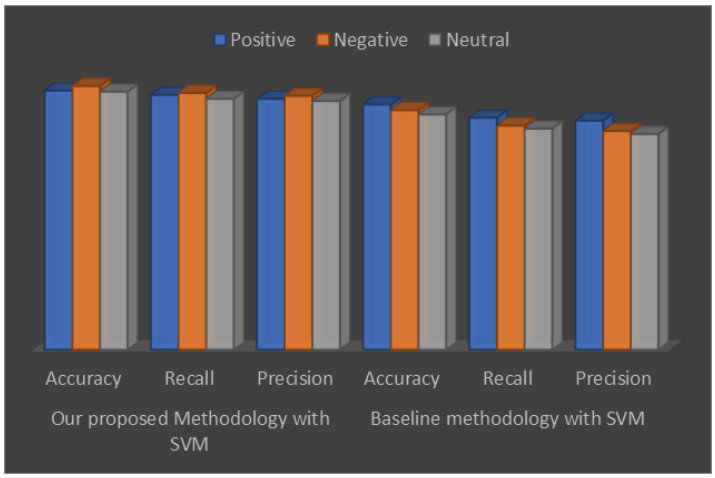


Graph 3: Overall comparison of classifier's accuracy

Then final comparison was our proposed methodology with SVM along with baseline methodology with SVM. Hence, we take SVM classifier and compare with base line SVM Classifier. Then we identified that our proposed methodology with SVM classifier analyses the sentiment and outperforms well when compared to traditional SVM classifier and the corresponding table and comparison are given as follow

	Our proposed Methodology with SVM			**Baseline methodology with SVM**		
	Accuracy	**Recall**	**Precision**	**Accuracy**	**Recall**	**Precision**
Positive	**96.5**	94.9	93.4	91.3	86.4	85.3
Negative	**98.2**	95.6	94.5	89.2	83.4	81.4
Neutral	**96.1**	93.4	92.6	87.6	82.3	80.3

Table 7: Comparison of proposed Methodology with SVM along with baseline methodology with SVM



Graph 4: Comparison of proposed Methodology with SVM along with baseline methodology with SVM

6. Conclusion & Future work

Usually, analysing some acute disease-affected people mindset is a much challenging role. Also, the dataset must be very accurate. Hence, we collected datasets from three different environments,

including a review from social media, a critical review from Twitter, and abstracts of medical study from the wall street journal. Then we proposed a practical framework that differs from the traditional approach, which includes four crucial techniques, i.e., Enhanced pre-processing, N-gram tokenization, assigning polarity score, and topic modelling with Latent Dirichlet Allocation. Finally, for evaluation purposes, we used a significant classifier that is prominently used in medical applications, including SVM, NB, DT, KNN, and RF. Out of this classifier, SVM outperforms well, and it got better accuracy of **98.2%**. Also, we compared our proposed practical framework with a baseline approach to prove our work analyses acute disease affected people emotion in an efficient way. We will incorporate this with a deep learning approach to processing some vast features of the dataset in future work.

Funding Information

No Funding

Conflict of Interests

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability statement:

Data sharing not applicable to this article as no datasets were generated or analysed during the Current study

Code Availability

Not Applicable.

Reference

[1] Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform. 2005;6(3):239–251.

[2] Jiang M, Chen Y, Liu M, Trent Rosenbloom S, Mani S, Denny JC, Hua X. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc. 2011;18(5):601–606.

[3] Cambria E. An introduction to concept-level sentiment analysis. In: Mexican international conference on artificial intelligence, pp 478–483. Springer. 2013.

[4] Cambria E. Affective computing and sentiment analysis. IEEE Intell Syst. 2016;31(2):102–107.

[5] Cambria E, Jie F, Bisio F, Poria S. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In: AAAI, pp 508–514. 2015.

[6] Swaminathan R, Sharma A, Yang H. Opinion mining for biomedical text data: Feature space design and feature selection. In: The 9th international workshop on data mining in bioinformatics, BIOKDD. 2010.

[7] Mondal A, Chaturvedi I, Das D, Bajpai R, Bandyopadhyay S. Lexical resource for medical events: A polarity based approach. In: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp 1302–1309. IEEE. 2015.

[8] Mondal A, Das D, Cambria E, Bandyopadhyay S. Wme: Sense, polarity and affinity based concept resource for medical events. In: Proceedings of the 8th global wordnet conference, pp 242–246. 2016.

[9] Mondal A, Satapathy R, Das D, Bandyopadhyay S. A hybrid approach based sentiment extraction from medical context. In: 4th workshop on sentiment analysis where ai meets psychology (SAAIP 2016), IJCAI 2016 Workshop, July 10, Hilton, New York City, USA. 2016.

[10] Basili R, Pazienza MT, Vindigni M. Corpus-driven unsupervised learning of verb subcategorization frames. In: Congress of the Italian Association for Artificial Intelligence, pp 159–170. Springer. 1997.

[11] Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc. 2007;14(3):304–311.

[12] Morante R, Liekens A, Daelemans W. Learning the scope of negation in biomedical texts. In: Proceedings of the conference on empirical methods in natural language processing, pp 715–724. Association for Computational Linguistics. 2008.

[13] Jacob SG, Geetha Ramani R. Discovery of knowledge patterns in clinical data through data mining algorithms: multi-class categorization of breast tissue data. Int J Comput Appl. 2011;32(7):46–53

[14] Ficek M, Kencl L (2012) Inter-call mobility model: a spatio-temporal refnement of call data records using a gaussian mixture model. In: 2012 Proceedings IEEE INFOCOM. IEEE, pp 469–477. https ://doi.org/10.1109/infcom.2012.6195786

[15] Liang J, Liu P, Tan J, Bai S (2014) Sentiment classifcation based on AS-LDA model. Proc Comput Sci 31:511–516. https://doi. org/10.1016/j.procs.2014.05.296

[16] Baltas ABAK, Tsakalidis AK (2017) Algorithmic aspects of cloud computing. In: Lecture Notes in Computer Science, vol 10230. Springer, Berlin, pp 15–25

[17] Oneto L, Bisio F, Cambria E, Anguita D (2016) Statistical learning theory and ELM for big social data analysis. IEEE Comput Intell Mag 11(3): https://doi.org/10.1109/MCI.2016.25725

[18] Chen J, Pan X, Monga R, Bengio S, Jozefowicz R (2016) Revisiting distributed synchronous SGD. arXiv preprint arXiv:1604.00981.

[19] Nodarakis N, Sioutas S, Tsakalidis AK, Tzimas G (2016) large scale sentiment analysis on twitter with spark. In: EDBT/ICDT workshops, pp 1–8

[20] Du J, Xu J, Song H, Liu X, Tao C (2017) Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. J Biomed Semant 8(1):1–7. https://doi. org/10.1186/s13326-017-0120-6

[21] Denecke K, Nejdl W (2009) How valuable is medical social media data? Content analysis of the medical web. Inf Sci 179(12):1870– 1880. https://doi.org/10.1016/j.ins.2009.01.025 (Elsevier Inc)

[22] Xia L, Gentile AL, Munro J, Iria J (2009) Improving patient opinion mining through multi-step classifcation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artifcial Intelligence and Lecture Notes in Bioinformatics), 5729 LNAI, pp 70–76. https://doi.org/10.1007/978-3-642-04208-9_13

[23] Cambria E, Benson T, Eckl C, Hussain A (2012) Sentic PROMs: application of sentic computing to the development of a novel unifed framework for measuring health-care quality. Expert Syst Appl 39(12):10533–10543. https://doi.org/10.1016/j.eswa.2012.02.120

[24] De la Torre-Díez I, Díaz-Pernas FJ, Antón-Rodríguez M (2012) A content analysis of chronic diseases social groups on facebook and twitter. Telemed e-Health 18(6):404–408. https://doi.org/10.1089/ tmj.2011.0227

[25] Portier K, Greer GE, Rokach L, Ofek N, Wang Y, Biyani P, Yu M, Banerjee S, Zhao K, Mitra P, Yen J (2013) Understanding topics and sentiment in an online cancer survivor community. J Natl Cancer Inst Monogr 47:195–198. https://doi.org/10.1093/jncim onographs/lgt025

[26] Crannell WC, Clark E, Jones C, James TA, Moore J (2016) A patternmatched Twitter analysis of US cancer-patient sentiments. J Surg Res 206(2):536–542. https://doi.org/10.1016/j.jss.2016.06.050

[27] Chen Z, Zeng DD (2017) Mining online e-liquid reviews for opinion polarities about e-liquid features. BMC Public Health 17(1):1–7. https://doi.org/10.1186/s12889-017-4533-z

[28] Ozcift A, Gulten A (2011) Classifer ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed 104(3):443–451. https://doi.org/10.1016/j.cmpb.2011.03.018

[29] Chen M, Hao Y, Hwang K, Wang L, Wang L (2017a) Disease prediction by machine learning over big data from healthcare communities. IEEE Access 5(c):8869–8879. https://doi.org/10.1109/acces s.2017.2694446

[30] Chen T, Xu R, He Y, Wang X (2017b) Improving sentiment analysis via sentence type classifcation using BiLSTM-CRF and CNN. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2016.10.065

[31] Lin F, Xiahou J, Xu Z (2016) TCM clinic records data mining approaches based on weighted-LDA and multi-relationship LDA model. Multimed Tools Appl 75(22):14203–14232. https://doi. org/10.1007/s11042-016-3363-9

[32] Jonnalagadda S, Peeler R, Topham P (2012) Discovering opinion leaders for medical topics using news articles. J Biomed Semant 3(1):2 Kim E, Han JY, Moon TJ, Shaw B, Shah DV, McTavish FM, Gustafson DH (2012) The process and efect of supportive message expression and reception in online breast cancer support groups. PsychoOncology 21(5):531–540. https://doi.org/10.1002/pon.1942

[33] Lu Y (2013) Automatic topic identifcation of health-related messages in online health community using text classifcation. SpringerPlus 2(1):1–7. https://doi.org/10.1186/2193-1801-2-309

[34] Manogaran G, Varatharajan R, Priyan MK (2018) Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. Multimed Tools Appl 77(4):4379–4399. https://doi.org/10.1007/s1104 2-017-5515-y

[35] Minarro-Gimenez JA, Marin-Alonso O, Samwald M (2014) Exploring the application of deep learning techniques on medical text corpora. Stud Health Technol Inform 205:584–588. https://doi. org/10.3233/978-1-61499-432-9-584

[36] TH M, Sahu S, Anand A (2015) Evaluating distributed word representations for capturing semantics of biomedical concepts. In: Proceedings of BioNLP 15, (Ml), pp 158–163. https://doi. org/10.18653/v1/w15-3820

[37] Chiu B, Crichton G, Korhonen A, Pyysalo S (2016) How to train good word embeddings for biomedical NLP. In: Proceedings of the 15th workshop on biomedical natural language processing, pp 166– 174. https://doi.org/10.18653/v1/w16-2922

[38] Spinczyk D, Nabrdalik K, Rojewska K (2018) Computer aided sentiment analysis of anorexia nervosa patients' vocabulary. BioMed Eng Online BioMed Cent. https://doi.org/10.1186/s1293 8-018-0451-2

[39] Timusk T, Holmes CC, Reichardt W (1995) C-axis properties of 123, like Lanl-Cm95. Anharmonic Prop High-T_c Cuprates 49:171

[40] Aisopos F, Papadakis G, Varvarigou T (2011) Sentiment analysis of social media content using N-Gram graphs. In: Proceedings of the 3rd ACM SIGMM international workshop on Social media— WSM'11, p 9. https://doi.org/10.1145/2072609.2072614

[41] 41. Dey A, Jenamani M, Thakkar JJ (2018) Senti-N-Gram: an n-gram lexicon for sentiment analysis. Expert Syst Appl 103:92–105. https:// doi.org/10.1016/j.eswa.2018.03.004 (Elsevier Ltd)

[42] Vittayakorn S, Umeda T, Murasaki K, Sudo K, Okatani T, Yamaguchi K (2016) Automatic attribute discovery with neural activations, Lecture Notes in Computer Science (including subseries Lecture Notes in Artifcial Intelligence and Lecture Notes in Bioinformatics), 9908 LNCS, pp 252–268. https://doi.org/10.1007/978- 3-319-46493-0_16

[43] Miura Y, Hattori K, Ohkuma T, Masuichi H (2013) Topic modeling with sentiment clues and relaxed labeling schema. In: Proceedings of the 3rd workshop on sentiment analysis where AI meets psychology, pp 6–14.

[44] Sarker A, Molla-Aliod D, Paris C, et al. Outcome polarity ´ identification of medical papers, pp 105–114. 2011.

[45] Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, Wahner-Roedler DL. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak. 2005; 5(1):13.

[46] Goldin I, Chapman WW. Learning to detect negation with 'not'in medical texts. In: Proc workshop on text analysis and search for bioinformatics, ACM SIGIR. 2003

[47] Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc. 2007;14(3):304–311

[48] Bashri MFA, Kusumaningrum R (2017) Sentiment analysis using Latent Dirichlet allocation and topic polarity wordcloud visualization. In: 2017 5th international conference on information and communication technology, ICoIC7 2017, 0(c), pp 4–8. https:// doi.org/10.1109/icoict.2017.8074651

[49] Yu R, Li A, Morariu VI, Davis LS (2017) Visual relationship detection with internal and external linguistic knowledge distillation. In: Proceedings of the IEEE international conference on computer vision, 2017-Octob(1), pp 1068–1076. https://doi.org/10.1109/ iccv.2017.121