Journal of Ambient Intelligence and Humanized Computing (2023) 14:5309–5325
https://doi.org/10.1007/s12652-019-01399-8

ORIGINAL RESEARCH

Sentiment analysis and text categorization of cancer medical records with LSTM

Deepak Chowdary Edara[1] · Lakshmi Prasanna Vanukuri[1] · Venkatramaphanikumar Sistla[1] · Venkata Krishna Kishore Kolli[1]

Received: 7 February 2019 / Accepted: 9 July 2019 / Published online: 16 July 2019

Abstract
Cancer is one among leading diseases, which affects millions of people and families around the world. Monitoring the mood of such cancer affected people plays a vital part in their treatment. In recent days, social media provides a platform for many people to share their experiences about the cancer through various blogs and communities. In this study, we intended to analyse moods of various cancer affected patients by collecting tweets from different online cancer supported communities. We employed several text mining and machine learning strategies to perform sentiment analysis on a distributed framework and developed a model for easier and faster analysis. The proposed distributed framework with long short-term memory (LSTM) neural network is an alternative to the conventional sentiment analysis approaches in analysing large volumes of data in a potential flow. The effectiveness of proposed framework was evaluated on the proposed dataset (corpus-1) and other two benchmark datasets like Health news Tweets (corpus-2) and Medical abstracts (corpus-3). The performance of each text mining and classification method was separately evaluated on three datasets and compared to each other. The results proved that the proposed approach performed better among the other methods in terms of both accuracy and execution time.

Keywords Long short-term memory · Medical informatics · Cancer · Sentiment analysis · Principal component analysis · Latent Dirichlet Allocation

1 Introduction

As per Worldwide Cancer Statistics around 14.1 million people in this world are affected with cancer. In 2018, about 9.6 million deaths are estimated due to cancer and within that approximately 70% of deaths were caused in low-and-middle income countries. In the process of diagnosis, it is vital to consider the mood or behaviour of cancer affected patients in their treatment. Now-a-days, many patients are sharing their health conditions indifferent online health community portals, supporting organizations, pharmaceutical companies and cancer centres through social media. This kind of activities increases the awareness to the people about cancer by expressing their views, opinions and giving support to other patients in the portals. Today's social media contains lot of cancer supported communities (Qiu et al. 2011; Whitten et al. 2002; Zhao et al. 2014) provide an open channel for people to explore about cancer related issues. Typically, people bonding with these type of interactions gives support which could play an important role to their survivorship (Shaw et al. 2000; Kim et al. 2012). Based on that, sentiment analysis or opinion mining (Fang and Zhan 2015; Pang and Lee 2008) methods can be helpful to track the feelings of the cancer patients by examining their attitudes and behaviour in the form of posts and comments. For the past decade, a lot of work has been carried on cancer and heart-diseases as they are growing consistently every day. Many of the studies determined that "text mining" is helpful in the development of medical research, particularly those diseases like heart-diseases (Torii et al. 2015) and cancer. This must be done because it is a highly tedious task to extract useful knowledge from large amount of unstructured data. In such case, "Big Data" processing proved to be useful and also it has gained much attention in recent years as it becomes very common requirement (Cheng and Lau 2015). Now-a-days, with the rapid increase in technology, data can be collected faster and more easily. This raises the problem of how to discover useful knowledge

✉ Venkatramaphanikumar Sistla
svrphanikumar@yahoo.com

[1] Vignan's Foundation for Science, Technology and Research, Guntur, Andhra Pradesh, India

from high volumes of data. The process of Big Data is to gather bulk amount of data from various resources and to organize it in a meaning full way. Dealing or analysing the large amount of data helps in discovery of useful knowledge and meaningful patterns. Thus, big data requires scalable and efficient solutions that helps users to reachable at all levels of knowledge without any issues.

For this study, authors used apache spark framework with various text mining and machine learning strategies to examine and determine the sentiments of cancer patients (Tonks and Smith 1996). This framework provides great knowledge and information from a ridge of text which can be widely used in the field of medical research. The main objectives of this work are: to analyse the posts of various cancer patients obtained from different online peer support groups. To understand and identify the methods which makes sentiment analysis task as more convenient in health care area, by studying literature in this field and the methods they used. Moreover, we also suggest an archetype design analysis of user sentiments and opinions from the large-scale unstructured data. We also proved that the proposed distributed framework is suitable for faster analysis and computing. The rest of this paper is organized as follows: Sect. 2 describes a detailed related work. Section 3 summarizes about proposed methodology. Section 4 describes the experimental results and finally conclusion is presented in Sect. 5.

2 Related work

Recent literature has mainly aimed to work on analysing how social media was influenced to help the public in sharing health information. In this regard, many studies uses distributed analytics (Ficek and Kencl 2012; Rahnama 2014) by employing text mining strategies plays a vital role in processing high volumes of unstructured data. Baltas and Tsakalidis (2017) performed twitter sentiment analysis using apache spark with binary and ternary classification. Oneto et al. (2016) proposed a conventional extreme learning machine (ELM) model using spark cluster. Chen et al. (2016) proposed a scalable deep learning framework in mobile bigdata analytics using apache spark. Those results are clearly evident that deep learning with spark achieves higher performance when compared to other spark models. Nodarakis et al. (2016) also performed sentiment analysis using spark framework on large scale data. Du et al. (2017) proposed an optimized machine learning system to extract sentiments from HPV vaccines related tweets. They manually annotated 6000 tweets and performed hierarchical classification with SVM model. The results show better performance of 0.6732 F-score when compared to other baseline models. Alike, general sentiment analysis approaches, medical sentiment analysis has become active research area. As an example, Denecke and Nejdl (2009) proposed a method to measure credibility and content quality in patient generated content using subjectivity words. They also developed a medical ontology to assess factual content in the medical texts that appears in social media. Generally, sentiment analysis is performed either by rule or machine learning based approaches. In terms of methods, majority of works are presented with machine learning methods rather than rule-based approaches. Xia et al. (2009) proposed a multi-step opinion classification model to determine polarity in patient data. Cambria et al. (2012) proposed a framework by integrating Sentic PROMs with emotion analysis methods to measure healthcare quality. De la Torre-Díez et al. (2012) attempted to characterize breast cancer, diabetes and colorectal cancer content from social media groups. People later turned to characterize relationships of cancer patients on Twitter (Murthy and Eldredge 2016). Portier et al. (2013) applied sentiment analysis techniques to detect negative emotions and unenthusiastic mood changes in a person based on interactions in online cancer communities. Crannell et al. (2016) explored a study on analysing sentiment of families who are psychologically influenced by the patient. Chen and Zeng (2017) analysed online e-liquid reviews by extracting e-liquid features. They performed sentiment analysis to classify the polarity of features which were obtained from large online e-liquid websites. Ozcift and Gulten (2011) explored a study on improving performance of machine learning algorithms in medical diagnosis. They combined one machine learning classifier with a CFS algorithm to evaluate the classification performance. The resulting model was assessed on the three medical datasets and produced an improved accuracy rate of 74.47%, 80.49% and 87.13% when compared with base classifiers.

Chen et al. (2017a, b) proposed a CNN-MDRP model to predict disease risk from structured and unstructured data. The experiment was carried on real-life hospital dataset and reaches 94.8% accuracy with a convergence speed when compared to other existing algorithms. Lu (2013) proposed a topic identification model based on text classification. The proposed model was evaluated by collecting data from online health communities using different feature sets and classification methods. They also performed feature-based classification with C4.5, SVM and Naïve Bayes. The experimental results showed that SVM outperformed with an improved classification results among other methods. Chen et al. (2017a, b) proposed a unique approach for improving sentence level sentiment analysis. The evaluation was performed on different sentence level sentiment analysis datasets in comparison of eleven approaches. The results show that their proposed approach outperforms with other existing methods. Lin et al. (2016) discussed on TCM clinic records and obtained a multi-relationship model by combining several features using weighted LDA topic model. The performance of the proposed model was improved with better classification rate and produced a novel support in TCM

clinical research. Jonnalagadda et al. (2012) explored a study on identifying opinion leaders from 147,528 obesity news articles. They prepared a corpus with 734,204 samples and achieved 88.5% efficiency. A novel deep learning model has been proposed by Manogaran et al. (2018) for heart disease diagnosis with multiple kernel learning. Minarro-Gimenez et al. (2014) applied neural language models to PubMed corpus for the first time. They aimed to work on word representations from the large amount of PubMed text articles using skip-grams. After, the interest is growing with neural language models CBOW and Skip-gram (Carod et al. 1997). They aimed to work on word representations from the large amount of PubMed text articles using skip-grams. Later, TH et al. (2015) performed skip-gram and CBOW on 1.25 million PubMed articles by assessing the word embeddings with word pairs. Chiu et al. (2016) discussed about training of good word embeddings for Biomedical NLP. They experimented on two different corpora proved that skip-gram model achieves better outcome than CBOW. Spinczyk et al. (2018) proposed a rule-based model for analysing sentiments from the patients suffering with anorexia nervosa. Using bag of words approach, the sentiment terms are identified from the documents which could help people to focus on specific topics during therapy.

2.1 Problem synopsis

To overcome the shortcomings found at literature, the proposed approach helps as a novel approach for discovering trust in a straightforward way. In this study, we tested our approach with various supervised and unsupervised algorithms for determining opinions from the health reviews. During opinion extraction, a lot of challenges were raised (Liang et al. 2014) and that were solved by integrating an generative statistical model known as Latent Dirichlet Allocation (LDA). The identified opinions were evaluated by reducing the inappropriate terms from the LDA model using various feature selection and reduction approaches. The proposed work is organized as follows:

Input A review corpus
Output A predictable model.

(i) Initially, a corpus with number of reviews was shaped by performing different pre-processing techniques.
(ii) Important features were extracted using N-gram tokenization method.
(iii) TF-IDF was computed for each term to discover opinion polarity.
(iv) A probabilistic LDA (P-LDA) model was employed for combining the review data to form distributed topics.
(v) Significant terms were selected by Chi square feature selector.
(vi) Optimal terms were extracted by handling curse of dimensionality using principal component analysis.
(vii) Those reduced number of terms were classified with different classification models.
(viii) Finally, a classifier is chosen among all the models as an accurate model based on efficiency and time complexity like evaluation metrics.

3 Methodology

3.1 Distributed computing framework

Map reduce (Ha et al. 2015) is a parallel and distributed paradigm, which empowers the processing of large scale datasets across Hadoop cluster (Madani et al. 2018). Basically, map reducer collects the input from Hadoop distributed file system (HDFS) and it comprises of two main tasks called mapper and reducer. Mapper is a base class which offers the projection of input data based on the input splits offered to the worker node and results an output with a key-value pair. The sort and shuffle stages produce data sorting based on the specified key input and creates an understandable format for the reducer. Further, reducer collects the inputs from intermediate data and performs the transformation for a given key value. In general, map reducer is not suitable for real time data processing because which requires shuffled data over the network. In the case of scalable datasets, mapper and reducer take long time in processing and having high latency. To overcome all these limitations, we performed sentiment analysis over apache spark framework as depicted in Fig. 1.

Apache spark is one of the fastest data processing frameworks and which is ten times faster than map reduce model and used to address the limitations of map reduce model. Spark does not allocate data to a disk at each iteration and it processes the data through memory until it reaches to its capacity. Once the disk capacity becomes full, then it pushes the data into the main storage. From the Fig. 1, spark driver master acts as a master node and several spark workers as worker nodes and theses worker nodes are handled by spark driver. The spark worker contains executors that relate to a spark to distribute the data in a cluster. Then, cluster manager looks after the responsibility of executing the tasks by instructing all the worker nodes in a cluster. API as a driver program enables the users to post the request and to get the reply in the form of a spark session or spark context. The spark session or spark context serves as a centralized part of communicating with all the spark workers. At the centre, spark works with resilient distributed datasets (RDD) to make the administrations like data collection, parallelism and fault node identification. In this work, the proposed approach employs data frames for operating and storing the

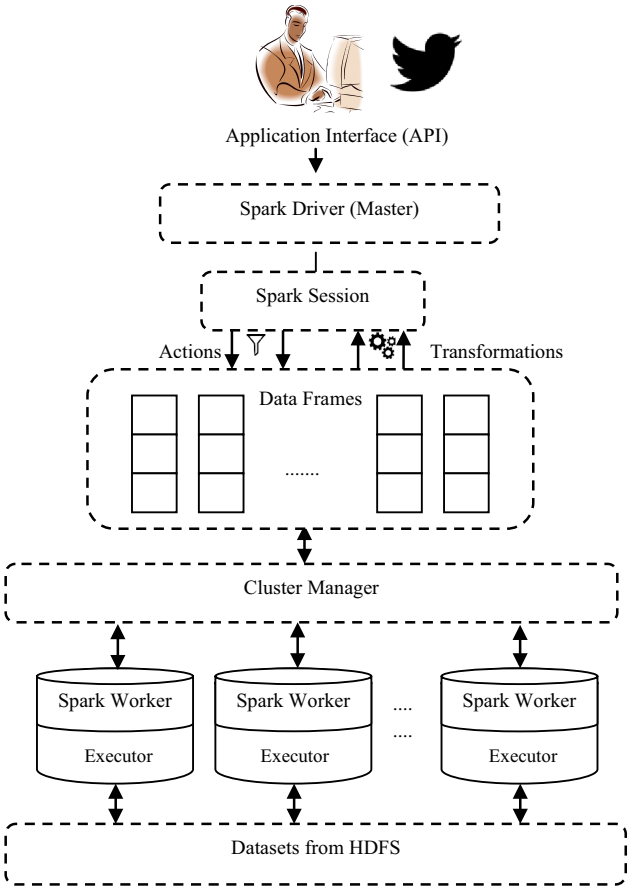

Fig. 1 Architecture of spark framework

data. These data frames contribute more in logical way, for operating tabular data. Spark model performs two essential operations called transformations and actions on the RDD's.

Transformation is applied on RDD to generate a new RDD from an existing one and then action is preformed to collect the data from those RDD's. The following Fig. 2 describes the overall implementation of the proposed system with Spark framework. Initially, this framework starts with data collection and performed pre-processing on the collected data. The following Tables 1 and 2 represents the detailed data statistics and various data pre-processing techniques that are used in this study. Later, feature extraction and feature selection are performed to extract relevant and more significant features from the pre-processed data. Further, classification was performed on the obtained features in the distributed computing model. The following sub-sections describes about the detailed explanation of the workflow.

3.2 Data collection

In this work, the efficiency of the proposed work is evaluated on three health and medical datasets. Initially, we have collected health tweets related to cancer from February 2, 2018 to October 2, 2018 using Twitter API[a]. We acquire 821,483 public tweets from 438,072 user's tweets with cancer related terms from various online cancer communities as represented in Fig. 3. The detailed description of this data is described in below sections. Similarly, we also utilized another benchmark Twitter dataset from UCI Machine Learning Repository[b]. This dataset contains health related tweets collected from various Twitter accounts like *reutershealth, kaiserhealthnews, bbchealth, NBChealth, nytimeshealth, everydayhealth, foxnewshealth, goodhealth, latimeshealth, msnhealthnews, cbchealth, wsjhealth, usnewshealth, cnnhealth, gdnhealthcare, and nprhealth* from August 2011 to December 2014. Furthermore, the third dataset is a larger dataset containing medical abstracts collected from Wall Street Journal[c].

1. https://developer.twitter.com/en/docs.html
2. https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter
3. https://sourceforge.net/projects/corpusredundanc/files/?source=navbar.

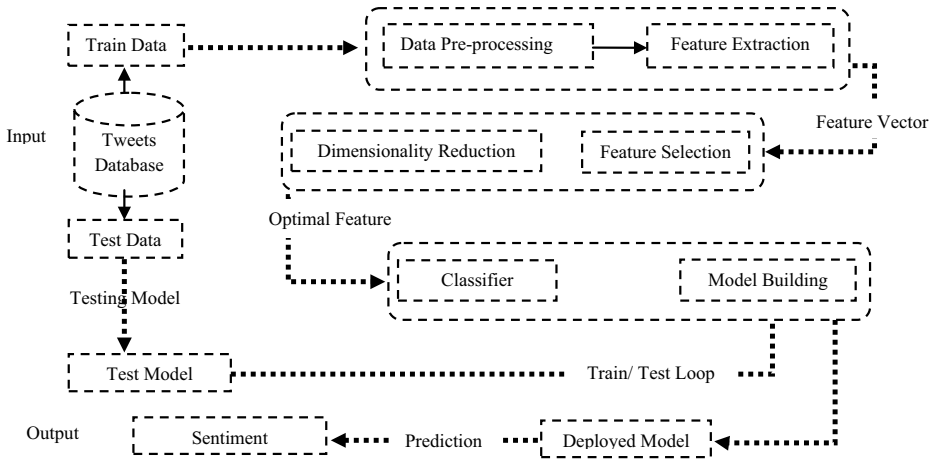

Fig. 2 Overview of proposed framework in spark cluster

Table 1 Statistics about the data

Characteristics	Count
Emoticons	64,072
URL	60,112
Tags	753,096
Unique words	780,422
Stop words	6,136,242
Positive words	783,221
Negative words	632,546
Neutral words	1,037,412
Total terms	10,247,123

Table 2 List of example emoticons

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

Fig. 3 Online cancer related communities in twitter social media

3.3 Data pre-processing

The real-world data collected from various data communities comprises lots of noise as well as it is required to be pre-processed for extracting relevant features, which are necessary for modelling. Mostly, the quality of corpus is unprocessed, and it is essential to pre-process the data before moving it into further phases of analysis. Initially, all the instances with emoticons and URL's were removed and then significance of each instance was identified. The hash tags were changed with defined words without hash mark.

Abbreviations misspelled, and slang words are processed by using regular expressions. In order to build meaningful corpus, all the less informative words are eliminated during this pre-processing phase. Pre-processing techniques like sanitization, stop-word removal and tokenization have been applied to diminish the corpus size by reducing unnecessary information. Through sanitization, all the numerical information is removed by transforming text into lower case from upper case. Lemmatization was performed to remove inflectional endings from the words. Stop-word removal is performed by eliminating all the English words which are not offering any necessary information. Additionally, we also developed three wordlists with 1200 extremely positive words, 1800 extremely negative words and 53 own stop-words to the existing positive, negative and stop-word lists. These wordlists were developed for understanding the influence of sentiment words in the data. The words in the wordlists are manually annotated with positive, extremely positive, negative and extremely negative sentiment labels by also including emoticons.

Table 3 Manually developed wordlist

Positive	Negative	Neutral
Health	Risk	Skin
Patient	Unsafe	Cancer
Favorite	Batvirus	Paracetamol
Protect	Wound	Syndrome
Natural	Misuse	Miscarriage
Prevent	Bacteria	Longevity
Treat	Bioterror	Symptom
Doctor	Damage	Suffocating
Clinic	Virus	Choice
Boost	Kill	Dilemma

Generally, emoticons also contain sentiments and are habitually typed in tweets. So, these types of emoticons were taken from Wikipedia (Szegedy et al. 2015; Devi et al. 2018) and developed an emoticons dictionary with 140 emoticons by labelling them as positive, negative or extremely positive, extremely negative. Table 1 describes about the preliminary statistics of the collected corpus and Table 2 shows the example of emoticon classification used in this work. Table 3 shows the example of developed word lists.

3.4 Feature extraction and selection

After pre-processing, N-gram tokenization (Timusk et al. 1995; Aisopos et al. 2011; Dey et al. 2018) was performed to extract the features by performing partitioning the text into a number of tokens. Basically, an N-gram is a set of occurring tokens in a frame and it is mostly used to predict the next tokens. Further, Sentence-level annotation and summarization is performed to extract the opinion words from the N-gram dictionary. This N-gram dictionary is manually annotated according to the pre-processed data and at

summarization, terms like "a", "and", "the", "there", etc… were eliminated from the review sentences because they won't contain any necessary information. Figure 4 shows the summary of each sentence with a frequency. After summarization and manual annotation, the score of each feature was computed to find number of positive, negative and neutral words from the summarized features is tabulated in Table 2. The score was computed separately for all positive words and negative words including their emoticons. Further, the opinion terms were identified and then processed into a feature vectorization and transformed into vectors. These feature vectors were evaluated based on Term Frequency and Inverse Document Frequency (TF-IDF) (Vittayakorn et al. 2016) measure by assigning weights to each feature vector. The significance of feature vector is measured using the Eq. (1).

$$TFIDF(t,d,D) = TF(t,d) \cdot IDF(t,D) \tag{1}$$

$$IDF(t,D) = log\frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

where 't' denotes the term that occurs number of times in a document 'd'. N represents the total number of documents in corpus and $|\{d \in D : t \in d\}|$ represents the term 't' appears in the total number of documents 'D'. Based on this calculation, all the weighted feature vectors are modelled into Latent Dirichlet Allocation (Miura et al. 2013) through a pipeline. LDA consists of a "Bayesian" optimizer to extract relevant features by transforming them into different number of topics. Each topic in LDA is considered as a "dimension"

Fig. 4 Example word cloud for summarized sentiment words of cancer related terms

and the following sub-section describes more about working of LDA model.

3.4.1 Topic modelling with Latent Dirichlet Allocation (LDA)

LDA (Bashri and Kusumaningrum 2017) is an unsupervised probabilistic method that models corpus into set of topics and then each topic is modelled as a distribution over words. In this work, LDA is employed on feature vectors to pursue the text in corpus followed by a hierarchical Bayesian approach. Assume there are 't' arrangements of topics in a corpus 'c', where corpus contains number of reviews 'r'. Topics in corpus can be considered as a polynomial probability distribution of feature vectors and each review in the corpus is arbitrarily produced by 'k' topics. The following Eq. (3) represents the feature extraction process with LDA. A feature vector 'f' is obtained by uniting the topics 't' for 'r' in corpus 'c' and sampled from each word distribution.

$$P(f_i) = \sum_{k=1}^{N} P\left(\frac{f_i}{t_i} = k\right) P(t_i = k) \tag{3}$$

where $P(t_i = k)$ represents the probability of topic k sampled for feature f_i for each review in corpus c. $P(f_i|t_i = k)$ represents the probability of f_i under topic k and N denotes the total number of topics. The above equation can be more simplified by assuming that $\phi^k = P(f_i|t_i = k)$ and refers to a multinomial distribution of feature vectors for topics k. $\theta^r = P(t)$ refers to a multinomial distribution of topics for review r. ϕ and θ are the estimated parameters defined for semantic representation of feature vectors and reviews. where N_r is denoted as number of features in a review 'r'. R represents the total number of reviews. α and β are known as hyper parameters for topic-word and review-topic Dirichlet distributions required in the process of corpus generation (Yu et al. 2017). θ is a review level variable sampled once per r and f is feature level variable sampled once for each word in r with N_r. At the same time, if there are any entailed features, it is hard to score them directly. Hence, Gibbs sampling is used to overcome such a limitation by acquiring the desired parameter values. It determines the number of topics from the review corpus and generates the probability distributions of the topics from reviews and features. Gibbs sampler uses a conditional probability and it is given as follows is Eq. (4),

$$P(t_i = k|t_{-i}, f_i, r_i, \ldots) \propto \frac{C^{FN}_{fij} + \beta}{\sum_{f=1}^{F} C^{FN}_{fj} + F \cdot \beta} \cdot \frac{C^{RN}_{r_{ij}} + \alpha}{\sum_{k=1}^{N} C^{RN}_{r_j} + N \cdot \alpha} \tag{4}$$

where $t_i = k$ represents the features f_i allocated to the topic k, and t_{-i} denotes the allocated topics to f_i. r_i represents ith number of reviews. R represents the number of reviews and

F denotes the number of features used in this study. C^{FN} and C^{RN} represents the topic-review matrix and C^{FN}_{fj} denotes the collection of features f_i as it is given to topic k without the current word f_i. $C^{RN}_{r_j}$ denotes the topic k given to some word in a review r without f_i. By sampling, the parameters θ and ϕ are obtained as following,

$$\hat{\phi}_i^{(k)} = \frac{C^{FN}_{f_{ij}} + \beta}{\sum_{f=1}^{F} C^{FN}_{f_j} + F \cdot \beta} \tag{5}$$

$$\hat{\theta}_j^{(r)} = \frac{C^{RN}_{r_{ij}} + \alpha}{\sum_{k=1}^{N} C^{RN}_{r_j} + N \cdot \alpha} \tag{6}$$

In this proposed work, LDA is utilized to discover the text from the review corpus and fuse them into latent topics. The Table 4 represents an example of probabilistic LDA topics on the review corpus. We have modelled the number of topics as $K = 100, 200, 300, 400$ and 500 on the review corpus. The important features were identified from the topics based on the probabilities correlated with each feature. Later, the implementation of feature selection is presented in below section to handle curse of dimensionality problem.

3.4.2 Dimensionality reduction and feature selection with Chi square and PCA

Chi square selector (Meesad et al. 2011) and principal component analysis (Underhill et al. 2007; Vinodhini and Chandrasekaran 2014, 2015) are used to extract relevant features from the LDA topic modelling for use in model construction. LDA yields vast number of dimensions (topics), and it is highly required to employ a feature selection model to handle the curse of dimensionality. These both models reduce the feature space, which can also improve the speed and learning attitude. In this paper, Chi square selector is used to extract highly relevant features from all the dimensions and then PCA is applied to reduce the no. of dimensions with a higher degree. To obtain principal components (uncorrelated variables), PCA is applied to compute orthogonal transformations of the variables that constitute the dimensions of the existing features. In general, PCA deals with variability mostly but not correlation, so Chi square is used to find out the variables with a high degree of correlation.

3.5 Text categorization and sentiment analysis with LSTM

In this work, authors explored five different types of classifiers available in Spark ML Lib platform to perform sentiment analysis. Multinomial Logistic Regression (MLR), Multinomial Naive Bayes (NB), Linear Support Vector Machine (LSVC), Multilayer Perceptron (MLP), and Decision Tree (DT) were used for classification. Long-Short Term Memory (LSTM) classifier (Liang et al. 2017) is called on PySpark for classification of opinions in the proposed model. MLR (Hamdan et al. 2015) is a type of binary classifier that is used to model dichotomous outcome variables. It uses a logistic function to determine the correlation among the sample class and the extracted features from the input. This MLR method handles the multi-class problem by fitting (N-1) independent binary logistic classifier model. At the same time, it arbitrarily selects one target class as a reference class and fits (N-1) regression models that compare each of the remaining classes to the reference class. The limitation with this MLR model is that it cannot handle data with large number of target classes. Additionally, it requires a larger dataset to obtain better performance. NB is a well-known binary classifier (Brody and Davidson 1998) and it assumes that for any given label b, the relationship among a conditionally independent feature a_i can be defined as Eqs. (7) and (8):

$$P\left(b|a_1, \ldots, a_\omega\right) \propto P(b) \prod_{i=1}^{\omega} P(a_i|b) \tag{7}$$

Table 4 An example LDA model with k number of topics

Topic 100			Topic 300			Topic 500	
Content	Probability		Content	Probability		Content	Probability
Cancer	0.036472	...	Health	0.019637	...	Child	0.019242
Ovarian	0.018138		Colorectal	0.017241		Lymphoma	0.018682
Melanoma	0.018138		Leukemia	0.035714		Cells	0.018068
Tumor	0.018068		Smoking	0.032684		Drug	0.018682
Doctor	0.017246		Heart	0.032684		Kidney	0.018068
Prostrate	0.016907		Malignant	0.018868		Gastric	0.018068
Obesity	0.019242		Sarcoma	0.016173		Thyroid	0.017241
Brain	0.016408		DNA	0.018682		Gland	0.018138
Diagnosis	0.018868		Germs	0.018068		Syndrome	0.016667
Osteosarcoma	0.016962		Risk	0.016627		Transmission	0.017246

$$P(b)\prod_{i=1}^{n}P(a_i|b) \rightarrow \hat{b} = argmax_b P(b)\prod_{i=1}^{\omega}P(a_i|b). \quad (8)$$

In the above Eqs. (7) and (8), ω denotes the feature count in a review with positive or negative label b. But in this paper, we have performed Multinomial Naïve Bayes (MNB) (Vittayakorn et al. 2016; Szegedy et al. 2015; Yu et al. 2017). Multinomial Naive Bayes is a probabilistic learning method used for efficient document classification. Initially, the probability of each class is computed with following equation:

$$P(C) = \frac{T_C}{T_R} \quad (9)$$

where T_R. represents the number of review(s) labeled with class 'C' and T_R represents the number of reviews given for training. Then, the probability of review to each class is computed with:

$$P(C|R) = P(C)\prod_{i=1}^{m}P(a_i \in R|C). \quad (10)$$

The above Eq. (10) shows the output of high probability class will be assigned as the review 'R' class. LSVC using linear kernel function supports only binary classification (Esuli and Sebastiani 2006). LSVC from Spark's ML classifier better suits to scalable datasets and it is widely used for classification in the field of machine learning that solves optimization problems. SVM finds an optimal hyperplane which acts as a separator between two classes and it also identifies an optimum marginacurve between two classes called as a maximum marginal classifier.

The larger the margin between the hyperplane provides a good generalization for classification of data. MLP is a feed-forward artificial neural network, maps set of inputs onto set of suitable output. Generally, MLP consists of multiple layers of nodes and each layer is interconnected to the next layer to form a network. The nodes of the hidden layer use an activation function based on sigmoid function, whereas output layer nodes use activation function based on Softmax function. MLP foraon md for network training containing multiple layers of computational units connected in a feed-forward way.

$$\text{Sid Function} : \sigma(z_i) = \frac{1}{(1+e^{-z_i})}. \quad (11)$$

$$\text{Softmax Function} : f(z_i) = \frac{e^{z_i}}{\left(\sum_{k=1} Ne^{z_k}\right)}. \quad (12)$$

'N' denotes the number of nodes in the output layers, and z_i is computed as $z_i = w_i x + b_i$ where b is the bias for each node and w_i is the weight of ith node.

LSTM is a special class of recurrent neural network (RNN), which have the ability of learning long-term dependencies (Soutner and Müller 2013) and overcomes the limitations of RNN. LSTM captures the input terms from the sentence in a distributed term representation form which is used to represent a term in vocabulary in the form of continuous values. Each term w in dictionary W is inserted into n-dimensional space ($L \in R^{n\times|W|}$). Typically, a LSTM network contains a cell state C_i and hidden state h_i and it also contains a set of recurrently connected memory cells which consists of each three multiplicative units: forget F_i unit, input I_i unit and output O_i unit with weights W_F, W_I, W_O and bias B_F, B_I, B_O respectively. These multiplicative units aid the LSTM memory cell to execute various operations like read, write, reset and allows memory cell to access and store the information over an epoch. "σ" is sigmoid function used in input, forget, and output units for generation of values in between 0 and 1. The following equations represent a LSTM memory cell which can be denoted as:

$$\mathrm{I_i} = \sigma\left(\mathrm{W_I}\left[\mathrm{x_i}, \mathrm{h_{i-1}}\right] + \mathrm{B_I}\right) \quad (13)$$

Input gate's function '$\mathrm{I_i}$' generates new memory state if the significance of the new word is considerable. Based on the input and past hidden states, input gate determines the worth of preserving the new word, and thus allows creation of new memory.

$$\mathrm{F_i} = \sigma\left(\mathrm{W_F}\left[\mathrm{x_i}, \mathrm{h_{i-1}}\right] + \mathrm{B_F}\right) \quad (14)$$

Forget gate '$\mathrm{F_i}$' is like the input gate but it determines whether the past memory cell is useful for the computation of the current memory cell or not. The forget gate acts on the input word and the past hidden state and produces $\mathrm{F_i}$.

$$\tilde{C}_\mathrm{i} = \tanh\left(\mathrm{W_C}\left[\mathrm{x_i}, \mathrm{h_{i-1}}\right] + \mathrm{B_C}\right) \quad (15)$$

where '$\tilde{C}_i$' is new memory which is based on aspects of new word '$\mathrm{x_i}$' and past hidden state '$\mathrm{h_{i-1}}$'

$$\mathrm{C_i} = \mathrm{F_i} \times \mathrm{C_{i-1}} + \mathrm{I_i} \times \tilde{\mathrm{C}}_\mathrm{i} \quad (16)$$

Based on outcome of forget gate '$\mathrm{F_i}$', it leaves out past memory '$\mathrm{C_{i-1}}$' in this stage. It is also takes outcome of input gate $\mathrm{I_i}$ and new memory $\tilde{C}_i$. Then the model sums these two results to produce the final memory '$\mathrm{C_i}$'.

$$\mathrm{O_i} = \sigma(\mathrm{W_O}[\mathrm{x_i}, \mathrm{h_{i-1}}] + \mathrm{B_O}) \quad (17)$$

$$\mathrm{h_i} = \mathrm{O_i} \times \tanh\left(\mathrm{C_i}\right) \quad (18)$$

Output gate '$\mathrm{O_i}$' determines when to output the value stored in the memory cell to the hidden layer. '$\mathrm{h_i}$' is new hidden state computed based on pointwise multiplying the output state and the new cell state.

4 Experimental results and discussions

This section presents a detailed study on the corpus and performance of the proposed method in terms of efficiency and computational complexity. Initially, a cluster is organized with 6 computing nodes configured with Linux Operating System; 8 GB RAM with 2.4 GHz processor and 1 TB hard disk. One among those six nodes is said to be as a Master Node and others were said to be Data nodes. The applications were built over Spark version 2.3.0 with Pyspark library installed using Python API on top of the Hadoop. The proposed approach has been executed on Spark data frames which takes the input in the form of tabular values.

4.1 Feature analysis and performance evaluation with running time on corpus 1, 2, 3

The corpus was duly processed in accordance with the following: (1) Significant terms associated with cancer feature must be present, (2) removal of duplicates in text and (3) removal of non-English characters other than emoticons from the text to evade difficulty in shaping multilingual tweets. Most of the tweets are undersized and have a bunch of odd words which makes the sentiment analysis task more complex. The obtained dataset was termed as corpus-1 and other benchmark datasets used in this study were termed as corpus-2 and corpus-3. The aim of using this corpus-1 is to determine the opinions of people on cancer through various online social media communities. During pre-processing phase, all the similar tweets were discarded from the corpora as they do not provide any useful information. The following Table 5 describes the statistics of terms extracted with different information retrieval techniques. After pre-processing, corpus -1 is comprised with 680,193 tweets with 7,652,217 terms and attained 7,173,144 terms by performing summarization. All the stop-words and sparse terms were removed during pre-processing and achieved 6,503,347 terms with TF-IDF. Further, the tokenized terms were considered as 577,318, 288,660 and 192,442 for unigrams, bigrams and trigrams by assigning polarity labels as positive, negative, and neutral.

Table 5 Statistics of terms extracted using different information retrieval techniques

Feature type	Corpus-1	Corpus-2	Corpus-3
Total terms	7,652,217	395,635	7,964,227
Summarized terms	7,173,144	285,987	7,236,137
Word2Vec	6,361,862	11,322	6,241,113
TF-IDF	6,503,347	11,974	6,592,318
Doc2Vec	6,257,942	10,486	6,112,579

Similarly, corpus-2 also consists of 58,927 health news, tweets with 395,635 terms. After pre-processing, 52,317 tweets were obtained with 285,987 terms after performing summarization. All the stop-words and sparse terms were removed during pre-processing and obtained 11,974 terms with TF-IDF. The tokenized terms were 1204 for unigrams, 603 for bigrams and 402 for Trigrams by assigning polarity labels as positive, negative, and neutral. Further, corpus-3 also contains a total of 34,611 medical documents with 7,964,227 terms. The obtained terms were 7,236,137 terms after performing tokenization and summarization. After removing stop-word and sparse terms, the terms obtained with TF-IDF were 6,592,318 and considered 601,498 for unigrams, 300,750 for bigrams and 200,450 for Trigrams. The sample unigram, bigram and trigrams extracted from a text review sample of corpus 1, 2 and 3 are presented in Table 6.

In this work, we aimed to work with bigrams because many of the previous works has been proved that working with bigrams extracts more sentiment information and achieves better results when compared to other n-grams (Ando et al. 2002; Barry 2017). The extracted bigrams are manually annotated by including medical terms to the available sentiment dictionaries such as SentiWordNet and labelled them as positive, negative, and neutral. Later, these reduced terms of various corpora were modelled into several topics using LDA as shown in below Figs. 5, 6 and 7.

Each topic in LDA results top terms from each tweet based on its frequency. As represented in below Figures, LDA results higher number of dimensions with methods such as Word2vec, TF-IDF and Doc2vec. As a part of topic modelling, we consider the size of N as 100, 200 and 300 topics. At this stage, the feature selection is highly required to obtain optimal number of dimensions to achieve better accuracy. In this regard, SVD and Chi square selector (CSS) were applied to extract more significant features and then PCA was employed to handle curse of dimensionality problem on the corpora. Finally, optimal topics that were obtained with CSS and PCA are used to perform classification considered for easier and faster computation. In the feature extraction phase, we have modelled LDA with N = 100, 200 and 300 topics based on the size of each extracted features.

Firstly, on corpus-1 singular value decomposition (SVD) was applied by considering 300 topics and extracted significant features with 230 topics. After that, Chi square is applied as a feature selector to obtain 185 relevant topics. Finally, PCA as a dimensionality reduction approach was incorporated and achieved an optimal 117 features. Similarly, the same topic modelling and feature selection was carried with 200 and 100 topics also. Further, the similar approach is applied on other two datasets and then those features were classified with various machine learning

Table 6 List of unigram, bigram and trigrams from a sample text review of Corpus 1, 2 and 3

Original review	Pre-processed review	Unigram	Bigram	Trigram
I attended a talk on cervical #cancer yesterday. Early diagnosis improves cancer outcomes by providing care at the earliest possible stage and is therefore an important to go for screening. It's treatable and has higher survival rate percentage when detected early.	Attended talk cervical cancer diagnosis improves cancer outcomes care earliest possible stage important screening treatable higher survival rate percentage detected	Attended; talk; cervical; cancer; diagnosis; improves; cancer; outcomes; care; possible; stage; important; screening; treatable; higher; survival; rate; percentage; detected	Attended talk; talk cervical; cervical cancer; cancer diagnosis; diagnosis improves; improves cancer; cancer outcomes; outcomes care; care possible; possible stage; stage important; important screening; screening treatable; treatable higher; higher survival; survival rate; rate percentage; percentage detected;	Attended talk cervical; talk cervical cancer; cervical cancer diagnosis; cancer diagnosis improves; diagnosis improves cancer; improves cancer outcomes; cancer outcomes care; outcomes care possible; care possible stage; possible stage important; stage important screening; important screening treatable; screening treatable higher; treatable higher survival; higher survival rate; survival rate percentage; rate percentage detected
Risk of colorectal cancer after a negative colonoscopy in low-to-moderate risk individuals: impact of a 10-year colonoscopy	Risk colorectal cancer negative colonoscopy low moderate risk individuals impact colonoscopy	Risk; colorectal; cancer; negative; colonoscopy; low; moderate; risk; individuals; impact; colonoscopy	Risk colorectal; colorectal cancer; cancer negative; negative colonoscopy; colonoscopy low; low moderate; moderate risk; risk individuals; individuals impact; impact colonoscopy	Risk colorectal cancer; colorectal cancer negative; cancer negative colonoscopy; negative colonoscopy low; colonoscopy low moderate; low moderate risk; moderate risk individuals; risk individuals impact; individuals impact colonoscopy
Air pollution? Obesity? Genetics? You can study them all via daylight saving time http://lat.ms/1A84HY3	Air pollution obesity genetics study daylight saving time	Air; pollution; obesity; genetics; study; daylight; saving; time	Air pollution; pollution obesity; obesity genetics; genetics study; study daylight; daylight saving; saving time;	Air pollution obesity; pollution obesity genetics; obesity genetics study; genetics study daylight; study daylight saving; daylight saving time

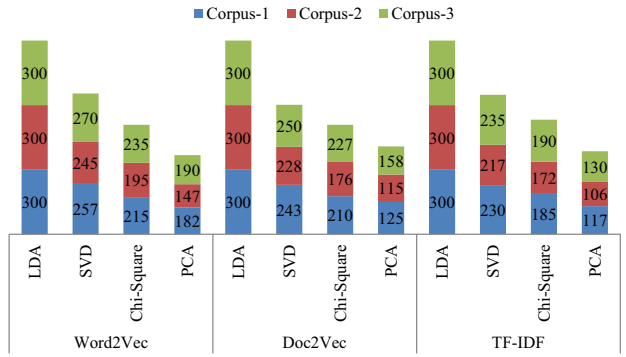

Fig. 5 No. of reduced topics with different feature extraction techniques when N = 300

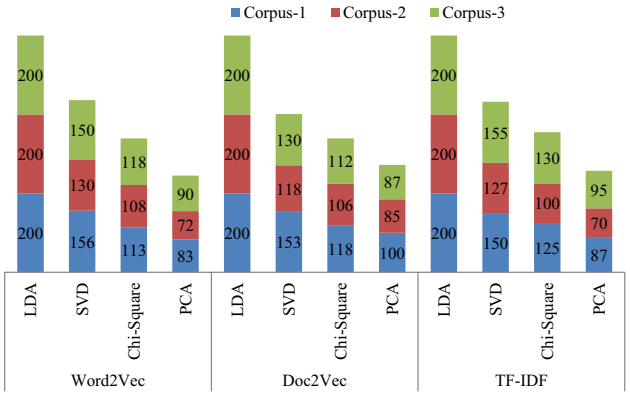

Fig. 6 No. of reduced topics with different feature extraction techniques when N = 200

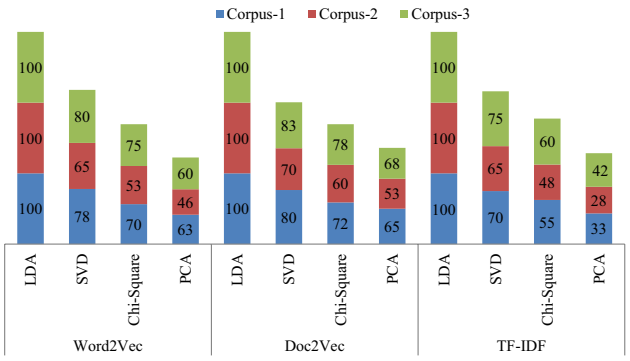

Fig. 7 No. of reduced topics with different feature extraction techniques when N = 100

classifiers available in MLlib and LSTM. The performance of each algorithm was assessed with metrics like accuracy and running time with tenfold cross validation. The trained LSTM model performs better among other classifiers. The following Tables 7 and 8 present the achieved results for all corpora MLR, NB, LSVC, DTree, MLP, and LSTM on both single node and multi-node cluster machines. From the results, it is shown that the proposed framework performed very efficient in most of the cases with an improved accuracy of 97.84% on corpus-1 and 88.37% on corpus-2 and 84.1% on corpus-3. Finally, the running time of all models with all feature extraction, feature selection (Yan et al. 2012) over single node and multi-node cluster were represented in Figs. 8 and 9.

4.2 Discussions and findings

This work extends the body of literature that emphasizes the significance of machine learning for publicly available large-scale datasets. Working with social media allows for examining the discussions happening outside of the public health space. The outcome of this work gives an example of utilizing various machine learning methods to estimate the gigantic social media landscape around cancer. We performed sentiment analysis of cancer patients by collecting 821,483 user tweets from various online cancer communities between February 2, 2018 to October 2, 2018 using Twitter API. Our analysis in this study also found that, majority of the people with cancer exhibits positive feelings regarding their support, treatment, and awareness openly on social media. However, our study also found some errors in the positive, negative and neutral categories. This determines the low accuracy obtained with some existing models. To examine errors in data, we analysed the misclassified samples in each corpus. We identified 3 major causes of errors in this study. They are sarcastic sentences, slang words, and word indistinctness. The first issue arises when there is a difficulty in finding the polarity. For example, if there are two sentences with different polarities then it is said to be a sarcastic sentence. At this state, it is a challenging task to solve such issues in opinion mining. The second issue is about making slang words which are commonly seen in social media. This issue affects the text because there are no spaces between words and gives a new meaning to the original text. Updating the existing polarity lexicon is only the solution to overcome this challenge. And such updating at every time might be a difficult task. The third problem shows the difficulties in assigning polarities to word. Depending on the context, a simple word in the text may contain many meanings. This would create some difficulties in detecting polarities from such words. To solve this issue, words connected with each other must be considered for successful opinion identification.

The framework presented in this study addresses those issues and can be applied to find similar knowledge about other public health-related topics. This study used manually annotated corpus to train the machine learning classifiers to analyse the sentiments from cancer related content on Twitter using spark framework. Of the examination of various feature extraction and feature selection techniques, bigrams

Table 7 Performance evaluation of different classifiers on single node with various feature selection techniques

Dataset	Approach	Word2Vec approach				Doc2Vec approach				TF-IDF approach			
		Without feature selection (LDA) (%)	With feature selection + feature reduction			Without feature selection (LDA) (%)	With feature selection + feature reduction			Without feature selection (LDA) (%)	With feature selection + feature reduction		
			SVD (%)	CSS (%)	PCA (%)		SVD (%)	CSS (%)	PCA (%)		SVD (%)	CSS (%)	PCA (%)
Corpus-1	MLR	58.33	61.27	64.52	67.64	55.12	58.14	60.23	63.08	63.64	66.43	74.12	79.63
	NB	62.24	66.03	66.48	71.24	57.58	60.46	62.24	66.52	65.32	69.12	77.36	85.21
	LSVC	60.16	65.27	67.12	72.13	58.72	61.82	64.34	67.18	66.26	71.08	78.23	86.54
	DTree	59.33	64.13	66.27	69.38	56.89	59.08	62.18	65.41	63.12	68.08	76.97	83.38
	MLP	56.58	60.12	63.36	66.16	54.47	57.71	58.36	61	61.82	65.52	73.32	81.65
	CNN	67.12	69.18	74.57	76.83	65.14	67.37	69.44	70.16	67.26	73	84.21	92.10
	LSTM	64.48	70.13	73.22	78.06	64.36	65.94	72	75.32	71.32	75.43	87.40	94.07
Corpus-2	MLR	56.10	60.32	62.36	65.33	52.33	57.24	59	62.87	60.08	64.07	68.56	74.16
	NB	59.26	63.12	64.45	70.62	60.74	62.16	64.18	66.18	64.10	68.26	74.86	82.09
	LSVC	60.37	64.86	65.16	72.87	59.12	62.06	64.36	67.42	65.77	69.83	74.10	81
	DTree	58.54	62.92	65.04	69.16	55.37	61.57	62.67	66.64	64.38	66.65	73.12	79.53
	MLP	55.86	58.07	62.36	65.12	52.58	57.64	59.08	63.12	60.54	63.28	67.92	71.33
	CNN	62.12	65.36	68.24	74.24	59.46	63.92	68	72.67	68.12	72.57	74.24	80.28
	LSTM	64.33	68.21	72.57	79.18	59.16	63.42	69.21	74.92	67.58	70.56	75.18	84.62
Corpus-3	MLR	55.48	59.33	60.18	65.68	51.14	53.57	54.89	59.54	59.13	63.10	68.54	74.64
	NB	57.14	61.14	63	68.87	53.56	56.15	58.47	62.31	63.42	67.18	73.96	78.72
	LSVC	58.37	62.16	65.12	69.12	57.17	59.31	59.32	63.45	62.43	65.89	72.45	77.10
	DTree	57.02	60.46	62.12	66.06	53.94	55.02	57.71	60.63	61.10	63.34	71.82	75.77
	MLP	53.86	56.27	58.92	63.48	52.31	56.12	54.32	58.14	57.64	58.43	68.12	73.12
	CNN	58.10	62.10	64.48	73.24	54.47	58.15	62.14	68.33	62.32	69.08	75.36	81.36
	LSTM	58	61.27	66.32	74.13	56.28	61.10	64.42	66.28	63.26	68.82	74.23	82.18

Table 8 Performance evaluation of different classifiers on a cluster with various feature selection techniques

Dataset	Approach	Word2Vec approach				Doc2Vec approach				TF-IDF approach			
		Without feature selection (LDA)	With feature selection + feature reduction			Without feature selection (LDA)	With feature selection + feature reduction			Without feature selection (LDA)	With feature selection + feature reduction		
			SVD	CSS	PCA		SVD	CSS	PCA		SVD	CSS	PCA
Corpus-1	MLR	58.60	62.52	64.98	69.10	56.08	60.12	63.14	66.23	64.15	68.31	75.18	80.66
	NB	62.76	66.97	68.18	73.16	59.86	62.86	64.82	67	67.26	71.47	77.82	86.07
	LSVC	62.24	66.10	69	74.28	58.25	62.10	65.10	67.98	69.31	74.61	79	86.78
	DTree	60.08	64.45	66.86	70.42	57	60.33	63	66.86	66.45	69.78	77.06	84.12
	MLP	56.58	60.87	63.96	68.10	54.12	58	61.24	64	63.34	68.44	73.69	82
	CNN	67.92	69.94	75.02	77	66	68.11	70	73.89	69.45	73.11	85.32	93.57
	LSTM	65.10	70.56	73.87	80.54	64.92	67.45	71.38	75.92	73.14	76.89	89.10	97.84
Corpus-2	MLR	56.27	61	63.92	66.74	53	58.12	60.47	65.84	62.06	65.14	68.33	75.23
	NB	60	64.02	67.23	71.22	61.46	63.74	66.38	69.17	67.44	70	75.04	84.34
	LSVC	61.07	65.18	67.18	73.56	59	62.86	64.02	67.56	67.31	70.94	74.67	84.63
	DTree	59.12	63.10	65.84	71.63	56.85	61.57	63	67.12	66.57	69.36	73.82	82
	MLP	56	58.47	63	67.48	53.06	59.32	62.08	66	61.36	64.77	68.16	73.98
	CNN	62.78	66.24	69.78	73.10	61.38	64.92	68.33	74.06	69.48	72.14	75.14	84.84
	LSTM	64.95	69	73.10	80	60.16	63.56	70	75.08	69.33	73.65	75.89	88.37
Corpus-3	MLR	55.10	59.82	62.36	66.32	52.22	56.17	59.29	62.27	61.14	65.16	69.73	75.12
	NB	57.94	62.08	64.11	68.27	53.56	57.92	59.28	63.08	65.97	68.37	74.33	80.33
	LSVC	58.82	62.90	65.89	68.92	57.04	60	62.74	65.86	63.89	66.92	75	79
	DTree	57.60	60.87	63.22	66.37	54.32	58.47	60.56	64.83	62.47	65.10	73.02	77.34
	MLP	54	56.47	59	63.88	52.96	57.66	61.92	63.14	59.66	62.38	68.54	74.09
	CNN	58.72	62.36	64.97	73	55	58.85	63.79	69	64.47	70	76.89	83.42
	LSTM	58.16	62.15	67	74.86	57.18	61.90	65	69.72	65.12	71.33	75.83	84.10

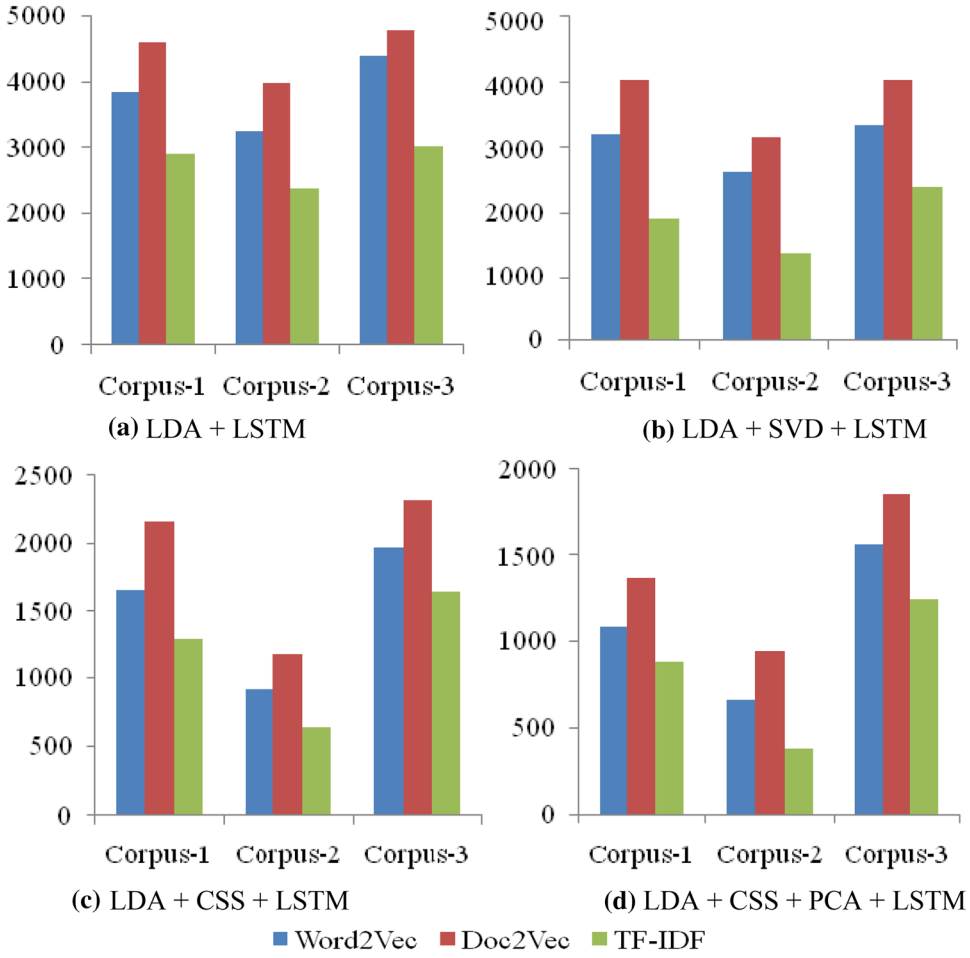

Fig. 8 Time complexities achieved on corpus-1, 2 and 3 using various feature extraction methods on a single node

with LDA, Chi square selector and PCA outperforms among other models like doc2vec, word2vec, and SVD. These were generally found to be successful for extracting more significant and optimal features. From the literature, manual annotation of sentiment data in medical field provides better results by constructing a quality dataset. Generally, social media platforms consist of 30% of non-relevant information which restricts the accuracy of real-time sentiment analysis. Automatic annotation requires considerable investment in the preparation of the lexicon and scripting of the automated tagging is complex task and results in loosing syntactic information and their relations. On the other hand, manual annotation overcomes the problem with minimal enhancements of the lexicon. Manual annotation is mainly limited in time taking for training data construction. The findings from above graphs and Tables assess the performances of classification methods evaluated on various corpora in terms of accuracy using proposed approach on both single node and cluster. Apache spark, a distributed computing framework supports data processing and querying on large scale datasets. It is found to be highly productive and fast in real-time data analytics. LSTM better handles time series data when compared with CNN and FCNN because it can make use of internal memory to process arbitrary input length of text sequences. Word embeddings will be clearly notified and memorized in LSTMs when compared to CNN. From the experimentation, it is clearly evidenced that the proposed work outperforms CNN on all three corpuses. The popular models such as NB, LSVC, DTree and MLP also show better rates of accuracy. These models do not attain extremely better accuracy but successful in terms of runtime. These obtained results prove that the proposed approach will be suitable for performing sentiment analysis in easier way in medical research field.

5 Conclusion and future scope

In this proposed work, we implemented a distributed framework to analyse mood of cancer affected patients from various online cancer supporting communities. The corpus was constructed by collecting patient reviews from various domains using Twitter API. It was manually annotated and well pre-processed to remove un-necessary information. Later, feature extraction followed by N-gram tokenization was employed to extract highly relevant features

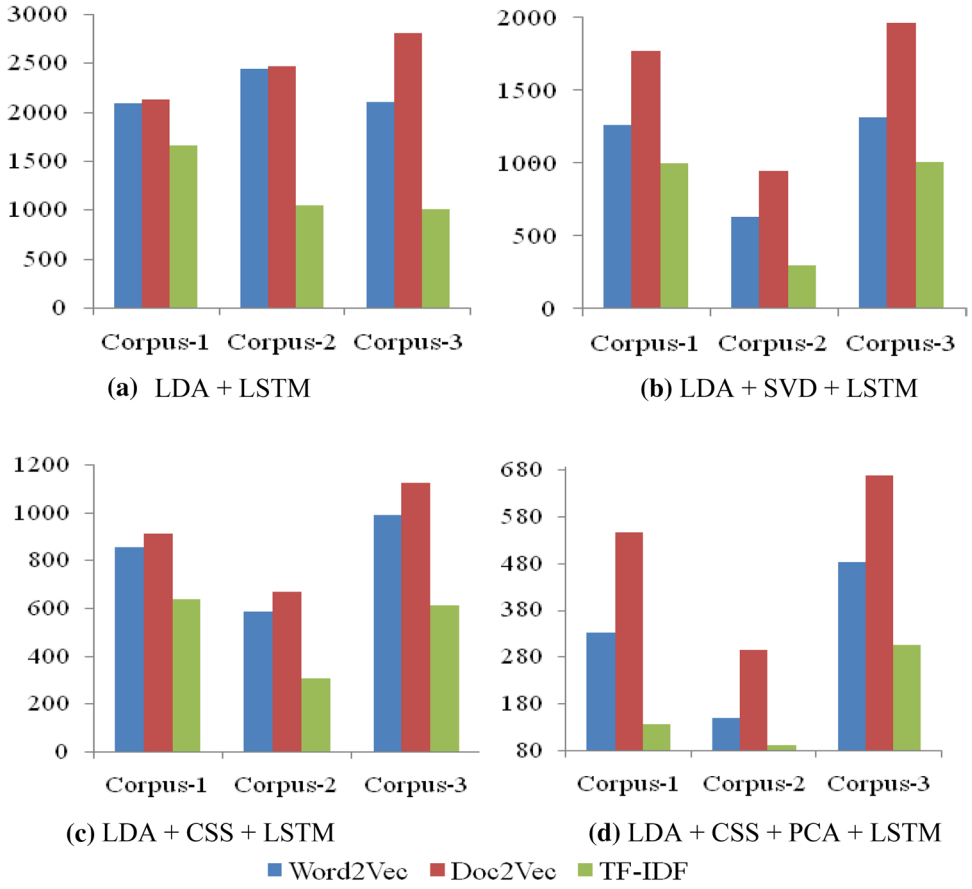

Fig. 9 Time complexities achieved on corpus-1, 2 and 3 using various feature extraction methods on multi-node cluster

using LDA topic modelling. Performance of the proposed work was evaluated and then compared with various classifiers such as MLR, NB, LSVC, MLP, DTree and LSTM. Based on the results, the performance evaluation of this proposed approach outperforms on both single and multi-node machines. Finally, we found that majority of the patients were expressed positive and some expressed negative and neutral about their disease. In immediate future, this proposed methodology can be extended with some potential feature extraction and feature selection techniques which work more efficiently on distributed environment. Furthermore, we will plan to propose an improved machine learning model that would be suitable for huge volumes of data at a faster rate. In this regard, we could also expect some potential extensions to this methodology for the development of sentiment analysis in health care field by providing valuable contributions for the researchers in future.

References

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

Ando Y, Terazaki H, Haraoka K, Tajiri T, Nakamura M, Obayashi K, Ishizaki T (2002) Presence of autoantibody against ATTR Val30Met after sequential liver transplantation. Transplantation 73(5):751–755. https://doi.org/10.1097/00007890-200203150-00016

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

Barry J (2017) Sentiment analysis of online reviews using bag-of-words and LSTM approaches. In: CEUR workshop proceedings, pp 272–274

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

Brody CM, Davidson N (eds) (1998) Professional development for cooperative learning: issues and approaches. Suny Press, New York

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

Carod FA, Cuadrado MP, González JG, Egido JH (1997) Autonomic disorder and sudden death in a patient with Wallenberg's syndrome. Neurología (Barcelona, Spain) 12(1):1–9. https://doi.org/10.1162/jmlr.2003.3.4-5.951

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

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

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/access.2017.2694446

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

Cheng OKM, Lau R (2015) Big data stream analytics for near real-time sentiment analysis. J Comput Commun 3(3):189–195. https://doi.org/10.4236/jcc.2015.35024

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

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

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

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)**

Devi KA, Edara DC, Sistla VPK, Kolli VKK (2018) Extended correlated principal component analysis with SVM-PUK in opinion mining. Turk J Electr Eng Comput Sci 26(5):2570–2582. https://doi.org/10.3906/elk-1704-178

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)**

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

Esuli A, Sebastiani F (2006) Determining term subjectivity and term orientation for opinion mining. In: Proceedings of the 11th meeting of the european chapter of the association for computational linguistics (EACL-2006), vol 2(1), pp 193–200. http://doi.org/10.1.1.60.8645

Fang X, Zhan J (2015) Sentiment analysis using product review data. J Big Data 2(1):5. https://doi.org/10.1186/s40537-015-0015-2

Ficek M, Kencl L (2012) Inter-call mobility model: a spatio-temporal refinement 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

Ha I, Back B, Ahn B (2015) MapReduce functions to analyze sentiment information from social big data. Int J Distrib Sens Netw. https://doi.org/10.1155/2015/417502

Hamdan H, Bellot P, Bechet F (2015) Lsislif: CRF and logistic regression for opinion target extraction and sentiment polarity analysis. In: Proceedings of the 9th international workshop on semantic evaluation, (SemEval), pp 753–758. https://doi.org/10.1016/j.crhy.2009.03.001

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 effect of supportive message expression and reception in online breast cancer support groups. Psycho-Oncology 21(5):531–540. https://doi.org/10.1002/pon.1942

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

Liang X, Lin L, Shen X, Feng J, Yan S, Xing EP (2017) Interpretable structure-evolving LSTM. In: Proceedings—30th IEEE conference on computer vision and pattern recognition, CVPR 2017, 2017-Janua, pp 2175–2184. https://doi.org/10.1109/cvpr.2017.234

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

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

Madani Y, Erritali M, Bengourram J (2018) Sentiment analysis using semantic similarity and Hadoop MapReduce. Knowl Inf Syst. https://doi.org/10.1007/s10115-018-1212-z **(Springer London)**

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/s11042-017-5515-y

Meesad P, Boonrawd P, Nuipian V (2011) A Chi square-test for word importance differentiation in text classification. Int Conf Inf Electron Eng 6:110–114. https://doi.org/10.1016/S0043-1354(01)00016-1

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

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

Murthy D, Eldredge M (2016) Who tweets about cancer? An analysis of cancer-related tweets in the USA. Digit Health 2:205520761665767. https://doi.org/10.1177/2055207616657670

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

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):45–55. https://doi.org/10.1109/MCI.2016.2572540

Ozcift A, Gulten A (2011) Classifier 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 **(Elsevier Ireland Ltd)**

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135. https://doi.org/10.1561/1500000011

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/jncimonographs/lgt025

Qiu B, Zhao K, Mitra P, Wu D, Caragea C, Yen J, Portier K (2011) Get online support, feel better-sentiment analysis and dynamics in an online cancer survivor community. In: Proceedings—2011 IEEE international conference on privacy, security, risk and trust and IEEE international conference on social computing, PASSAT/SocialCom 2011, pp 274–281. https://doi.org/10.1109/passat/socialcom.2011.127

Rahnama AHA (2014) Distributed real-time sentiment analysis for big data social streams. In: Proceedings—2014 international conference on control, decision and information technologies, CoDIT 2014, pp 789–794. https://doi.org/10.1109/codit.2014.6996998

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

Shaw BR, McTavish F, Hawkins R, Gustafson DH, Pingree S (2000) Experiences of women with breast cancer: exchanging social support over the CHESS computer network. J Health Commun 5(2):135–159. https://doi.org/10.1080/108107300406866

Soutner D, Müller L (2013) Application of LSTM neural networks in language modelling. In: Habernal I, Matoušek V (eds) Text, speech, and dialogue. TSD 2013, Lecture notes in computer science, vol 8082. Springer, Berlin

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/s12938-018-0451-2

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 07-12-June, pp 1–9. https://doi.org/10.1109/cvpr.2015.7298594

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

Tonks A, Smith R (1996) Information in practice. BMJ (Clin Res Ed.) 313(7055):438. https://doi.org/10.1136/bmj.313.7055.438

Torii M, Fan JW, Yang WL, Lee T, Wiley MT, Zisook DS, Huang Y (2015) Risk factor detection for heart disease by applying text analytics in electronic medical records. J Biomed Inform 58:S164–S170. https://doi.org/10.1016/j.jbi.2015.08.011 **(Elsevier Inc)**

Underhill DG, McDowell LK, Marchette DJ, Solka JL (2007) Enhancing text analysis via dimensionality reduction. In: 2007 IEEE international conference on information reuse and integration, IEEE IRI-2007, vol 21402(410), pp 348–353. https://doi.org/10.1109/iri.2007.4296645

Vinodhini G, Chandrasekaran RM (2014) Opinion mining using principal component analysis based ensemble model for e-commerce application. CSI Trans ICT 2(3):169–179. https://doi.org/10.1007/s40012-014-0055-3

Vinodhini G, Chandrasekaran RM (2015) Sentiment classification using principal component analysis based neural network model. In: 2014 International conference on information communication and embedded systems, ICICES 2014, vol 978, pp 1–6. https://doi.org/10.1109/icices.2014.7033961

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 Artificial Intelligence and Lecture Notes in Bioinformatics), 9908 LNCS, pp 252–268. https://doi.org/10.1007/978-3-319-46493-0_16

Whitten P, Mair F, Haycox A, May C, Williams L, Hellmich S (2002) Systematic review of cost effectiveness studies of telemedicine interventions. BMJ 324(7351):1434–1437

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

Yan X, Wu X, Kakadiaris IA, Shah SK (2012) To track or to detect? An ensemble framework for optimal selection. In: Fitzgibbon A, Lazebnik S, Perona P, Sato Y, Schmid C (eds) Computer vision—ECCV 2012. Lecture Notes in Computer Science, vol 75/76. Springer, Berlin

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

Zhao K, Yen J, Greer G, Qiu B, Mitra P, Portier K (2014) Finding influential users of online health communities: a new metric based on sentiment influence. J Am Med Inform Assoc JAMIA 21(e2):1. https://doi.org/10.1136/amiajnl-2013-002282