Nijhawan *et al. Journal of Big Data* (2022) 9:33
https://doi.org/10.1186/s40537-022-00575-6

Journal of Big Data

RESEARCH Open Access

Stress detection using natural language processing and machine learning over social interactions

Tanya Nijhawan[1], Girija Attigeri[2*] and T. Ananthakrishna[1]

*Correspondence:
ga.research10@gmail.com
[2] Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India
Full list of author information is available at the end of the article

Abstract

Cyberspace is a vast soapbox for people to post anything that they witness in their day-to-day lives. Social media content is mostly used for review, opinion, influence, or sentiment analysis. In this paper, we aim to extend sentiment and emotion analysis for detecting the stress of an individual based on the posts and comments shared by him/her on social networking platforms. We leverage large-scale datasets with tweets to accomplish sentiment analysis with the aid of machine learning algorithms and a deep learning model, BERT for sentiment classification. We also adopted Latent Dirichlet Allocation which is an unsupervised machine learning method for scanning a group of documents, recognizing the word and phrase patterns within them, and gathering word groups and alike expressions that most precisely illustrate a set of documents. This helps us to predict which topic is linked to the textual data. With the aid of these models, we will be able to detect the emotion of users online. Further, these emotions can be used to analyze stress or depression. In conclusion, the ML models and a BERT model have a very good detection rate. This research is useful for the well-being of one's mental health. The results are evaluated using various metrics at the macro and micro levels and indicate that the trained model detects the status of emotions based on social interactions.

Keywords: Decision tree, Latent Dirichlet Algorithm, Logistic regression, Machine learning, Natural Language Processing, Random forest, Sentiment analysis, Topic modelling

Introduction

Currently, social media plays the role of chief public opinion detector. We have over 4.2 billion active worldwide social media users. With the whirlwind expansion of Web 2.0, people have developed a liking to express their thoughts and approach over the Internet, which has consequently resulted in an increase of user-generated content and self-opinionated data. Social Media Analytics (SMA) is the process of collecting information on various social media platforms, websites and blogs and evaluating that, to successful business decisions. The use of social media has become quite commonplace in today's world. SMA is not only a collection of likes and comments shared by people but also a platform for many advertising brands. There are six types of social networks where

people connect and share their interests, opinions, experiences, and moments of life. Bookmarking sites allow users to have control over their resources. Social news: allows users to post news links and external articles, Media sharing: Share their videos and photographs, Microblogging: Allow users to write short written entries and Blogs and Forums: Allow users to produce focused content and then engage in conversations about it.

SMA is the ability to gather data from these resources and find meaning from them, make decisions and evaluate the performance of the decisions through social media. For this SMA uses the concepts such as social media intelligence, social media listening, social media monitoring, social competitive analysis, image analytics, sentiment analysis, customer sentiment analysis. Many applications include marketing and making extensive use of social data to make predictive decisions. Some of the methods are built to create a hypothesis, deep penetration of data, mapping events, etc. These calculations can also be done in services such as business, amendment, education, machine learning-based predictions, etc. Especially now, data is controlling marketing approaches and tactics. The propagation of data is only expected to rise as more people and businesses plan on dispensing data about themselves on social media. It is in this material that a business will end up learning more about their audience, specifically on sites like Twitter, Facebook, and Instagram. With these insightful analytics, a person fundamentally gains social media intelligence to inform future decisions and actions. Currently, SMA is being used for influence, review, and opinion mining. However, it can be employed in analyzing the emotional state of a person. These social factors are important indicators of mental health. However, how to be quantifying and analyzing social factors is challenging. The data is usually unstructured and huge, which needs the techniques like Big data, Machine Learning (ML), Natural Language Processing (NLP) to get inferences for stress or other mental health issues. There are studies to show that constructive SMA to measure and quantify the social interaction along with other health parameters are used in healthcare systems for stress/depression detection [1].

To perform the SMA data can be collected with the help of web scraping. Web scraping aids in extracting the underlying HTML code and, with it, the data deposited in a database. The scraper can then duplicate the complete website content elsewhere. Apart from this, with the help of applications like lucidya and trackmyhashtag, certain hashtags were tracked while creating the dataset. There are a lot of capable pre-trained language models which include the likes of ELMo and BERT. These models have specifically shown outstanding performance on aspect-based sentiment analysis problems [2]. The pre-trained language models have the advantage to learn universal language by pre-training on the vast unlabeled corpus to dodge overfitting on small-size data [3]. In this paper, we are using a proficient deep learning model titled BERT to resolve sentiment classification tasks. Experimentations have supported the claim that the BERT model outdoes other prevalent models for this task without a complex architecture. Hence, we use the BERT model to do a 5-class emotion classification. The emotions are joy, sadness, neutral, fear, and anger.

Topic modeling is described as one of the most efficacious methods for detecting useful unseen structures in a corpus. It can be defined as a method that locates a group of words i.e., the topic from a group of documents that represents the information in the group [4]. By leveraging the topic modeling results we can represent, measure, and model user behavior patterns on large-scale social networks and even use such social information for further research. With the edge of using ML algorithms, a pre-trained model, and a high accuracy rate, this model will give accurate and reliable results. The idea of this paper is to come up with a system that not only detects stress but also analyses the topic of discussion in a particular tweet. Along with sentiment analysis, this system will also accurately analyze and segregate the user's opinions on different topics. After carrying out in-depth studies on pertinent datasets we will attain crucial understandings of different correlations between social interactions and the tension/strain of the user.

The contributions of the paper are as follows:

- Binary classification of the sentiments behind the tweets
- Perform topic modeling with the help of LDA which takes into consideration the density of every topic and calculates a topic structure through an iteration process.
- Emotion classification using deep learning-based BERT model to detect stress.
- Develop a Django-based web application that receives inputs from a user and then accordingly generates a prediction.
- Develop a system in the form of a web portal that not only detects stress but also analyses the topic of discussion in a particular tweet.
- Accurately analyze and segregate the user's opinions on different topics.

Background and literature review

A lot of astounding contributions have been made in the field of sentiment analysis in the past few years. Initially, sentiment analysis was proposed for a simple binary classification that allocates evaluations to bipolar classes. Pak and Paroubek [5] came up with a model that categorizes the tweets into three classes. The three classes were objective, positive and negative. In their research model, they started by generating a collection of data by accumulating tweets. They took advantage of the Twitter API and would routinely interpret the tweets based on emoticons used. Using that twitter corpus, they were able to construct a sentiment classifier. This classifier was built on the technique—Naive Bayes where they used N-gram and POS-tags. They did face a drawback where the training set turned out to be less proficient since it only contained tweets having emoticons. The papers [6–10] discuss effective data pre-processing techniques for social media content, specifically tweets. As the data contains the words which are most often used in a sentence but do not contribute to the analysis, such as stop words, symbols, punctuation marks. Removing these and converting different forms of the words to the base from is an essential step.

Sentiment analysis

Agarwal et al. [11] proposed a 3-way model for categorizing sentiments in three classes. The classes were positive, negative, and neutral. Models such as the unigram model, a feature constructed upon the model, and a tree kernel-based were used for testing. In the case of the tree kernel-centered model, tweets were chosen to be represented in the form of a tree. While implementing a feature-centered model over 100 features were taken into consideration. However, in the case of the unigram model, there were about 10,000 features. They concluded that features that end up combining previous polarization of words with their parts-of-speech (pos) tags are the most substantial. In terms of the result, the tree kernel-based model ended up performing better than the other two models.

Certain challenges are made by a few researchers to classify public beliefs about movies, news, etc. from Twitter posts. V.M. Kiran et al. [12] utilized the data from other widely accessible databases like IMDB and Blippr after appropriate alterations to benefit Twitter sentiment analysis in the movie domain. Davidov et al. [13] projected a method to utilize Twitter user-defined hashtags in tweets as a classification of sentiment type using punctuation, single words, and patterns as disparate feature types. They are then combined into a single feature vector for the task of sentiment classification. They made use of the K-Nearest Neighbor approach to allocate sentiment labels by constructing a feature vector for each example in the training and test set. Tagging [14], in current times developed as a common way to sort out vast and vibrant web content. It usually refers to the act of correlating with or allocating some keyword or unit to a piece of data.

Tagging aids to depict an article and lets it be located again by perusing. Scholars have established diverse methods and procedures for tagging corpus for numerous uses. Xiance et al. [15] offered a flexible and practical technique for the process of the recommendation of tags. They demonstrated documents and tags by implementing the tag-LDA model. Krestel et al. [16] recommended a method to customize the process of recommendation by tag. She proposed a method that amalgamates a probabilistic method of tags from the source. In this case, the tags were extracted from the user. She examined basic language models. Additionally, she performed LDA experimentations on a real-world dataset. The dataset was crawled from a vast tagging system which displayed that personalization progresses the process of tag recommendation.

Pre-trained language models like ELMo [17], OpenAI GPT [18], and BERT [19] have proven to be extremely valuable. This has led to Natural Language Processing (NLP) passing into a new era. Transfer learning abilities permitted by pre-trained language models have helped a lot of researchers significantly. This has allowed the pre-trained model to play the role of the base, and this can be fine-tuned to respond to the NLP task. This process is better than performing the training of the model from the basics [20]. Zubair et al. [21] introduced a technique enhanced by lexicons. It was projected to be centered around a rule-based classification scheme. It was to be carried out by assimilating emojis, modifiers, and domain-specific terms to examine any thoughts published on social media. However, traditional methods emphasis on designing features has now reached its performance bottleneck [22]. On the other hand, pre-trained language models save a lot of time by achieving the same result quickly. They are easy to incorporate

Table 1 Comparison of different approaches in sentiment analysis

Researcher	Technique	Performances
Pak and Paroubek [5]	Naïve Bayes Sentiment classifier with multinomial features	High accuracy, Low decision value
Alec et al. [23]	Naïve Bayes classifier, Mutual information measure for feature selection	Accuracy: 81%
Balahur et al. [24]	WordNet-lexicon	Accuracy: 82% Improvement in the baseline 21%
Jonathon et Al. [25]	SVM, Naïve Bayes	Accuracy: 70%
Boiy et al. [26]	Integrated approach: ML, Information retrieval, NLP	Accuracy: 83% (English texts),70% accuracy (Dutch texts), 68% (French texts)
Li et al. [27]	Dependency- Sentiment, LDA, Markov chain	Accuracy: 70.7% with on tenfold cross-validation test set: 800 reviews

and there's not as much labeled data required. However, these techniques need to be incorporated for mental health prediction with social and other parameters (Table 1).

Stress/depression analysis

Arya and Mishra present a review of the application of machine learning in the health sector, their limitation, predictive analysis, and challenges in the area and need advanced research and technologies. The authors reviewed papers on mental stress detection using ML that used social networking sites, blogs, discussion forums, Questioner technique, clinical dataset, real-time data, Bio-signal technology (ECG, EEG), a wireless device, and suicidal tendency. The study shows the high potential of ML algorithms in mental health [28]. Aldarwish et al used machine learning algorithms SVM and Naïve- Bayesian for Predicting stress from UGC- User Generated Content in Social media sites (Facebook, Twitter, Live Journal) They used social interaction stress datasets based on mood and negativism and BDI- questionnaire having 6773 posts, 2073 depressed, 4700 non-depressed posts (textual). They achieved an accuracy of 57% from SVM and 63% from Naïve- Bayesian. They also emphasized stress detection using big data techniques [29].

Cho et al. presented the analysis of ML algorithms for diagnosing mental illness. They studied properties of mental health, techniques to identify, their limitations, and how ML algorithms are implemented. The authors considered SVM, GBM, KNN, Naïve Bayesian, KNN, Random Forest. The authors achieved 75% from the SVM classifier [30]. Reshma et.al proposed a Tensi Strength framework for detecting sentiment analysis on Twitter [31]. The authors considered SVM, NB, WSD, and n-gram techniques on large social media text for sentiment analysis and applied the Lexicon approach to detect stress and relaxation in large data set. The authors achieved 65% precision and 67% recall. Deshpande and Rao presented an emotion artificial intelligence technique to detect depression [32]. The authors collected 10,000 Tweet Using Twitter API. They applied SVM and Naïve Bayes machine learning algorithms and

achieved F1 scores of 79% and 83% respectively. Zucco et al. presented a preliminary design of an integrating Sentiment Analysis (SA) and Affective Computing (AC) methodologies for depression conditions monitoring [33]. The authors described SA and AC analysis pipelines. They also presented main challenges such as online learning and stream analytics for real-time processing in the design and implementation of such a system. These can be overcome by using big data technology. The authors have not presented the final system and the results testing and validation. The literature for stress detection shows that the models used for prediction need improvement. The mental health prediction and monitoring also need to be combined with other health parameters such as eating, sleeping, physiological and other factors.

Role of Big data in social media analytics

Sohangir et al. emphasized that deep Learning is a valuable tool for big data [34]. It can be used to extract remarkable information hidden in big data, considering social networks data. They considered the stock market as the domain. The authors aim to build deep Learning models to improve the performance of sentiment analysis using Stock Tweets. Authors applied neural network models such as LSTM, doc2vec, and CNN. They concluded that the deep Learning model can be used effectively for financial sentiment analysis with big data analytics.

Opinion mining is a significant area of NLP in big data utilizing data from social media. Applications are working on customer reviews, opinions for sentiment, influence analysis. Bandari and Bulusu used a clustering strategy for the classification of product reviews based on sentiment values [35]. Big data is filled with a volume of structured or unstructured data. The realization of online service depends on data from social media users, customers. Most of such data is voluminous and unstructured, hence requiring advanced techniques to handle big data such as Hadoop. Trupthi et al. proposed a feedback collection system based on structured query language [36]. The authors employed a decision tree for classifying reviews. A big data approach and machine learning algorithm are required for the efficient analysis of social media data. Hammou et al. proposed a neural network scheme in sentiment analysis; from this, they classified the customer emotions with high accuracy [37].

Considering the literature review, the focus of the current research is to leverage social media content by applying machine learning and deep learning techniques to predict the emotional state of a user. Further, use the analysis to detect stress. These models along with other health parameters can be used in assessing the mental health of a patient.

Research design and methodology

The research makes use of both secondary and primary data sources. It is a cross-sectional study and combination of quantitative and qualitative methodologies to know the impact of social and emotions associated with the social media data and usefulness of the same. The research aims at building models for sentiment and emotion detection which can be used for stress management, the models are also tested on primary data. The focus of the paper is identifying the sentiment or emotions of a user concerning diverse topics or domains using Latent Dirichlet Allocation (LDA). A hybrid machine

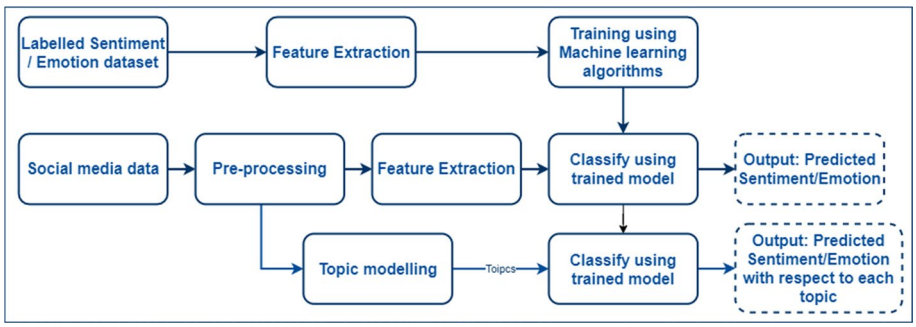


Fig. 1 Sentiment Analysis Methodology

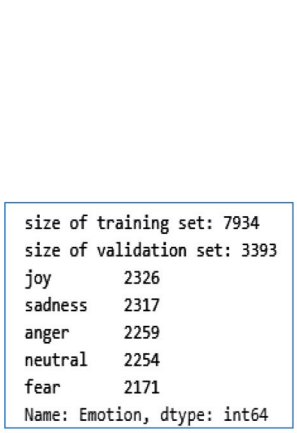
label	raw_tweet
0	love island Ã°ÂŸÂ'Â"
0	my fav actor #vijaysethupathi ! my fav actress @user ! my most fav director @user !! one film !! can't wait :)
0	whew Ã°ÂŸÃ˜… it's a productive and #friday!!!
0	@user she's finally here! @user
0	passed first year of uni #yay #love #pass #unistudent #photographystudent #ucs #soproud
0	this week is flying by #humpday - #wednesday #kamp #ucsdÃ¢Â€Â¦
0	@user modeling photoshoot this friday yay #model #me #follow #emo
0	you're surrounded by people who love you (even more than you deserve) and yet, why are so hateful?
0	feel like... Ã°ÂŸÂ˜ÂÃ°ÂŸÂÂ¶Ã°ÂŸÂ˜ÂŽ #dog #summer #hot #help #sun #day #more
1	@user omfg i'm offended! i'm a mailbox and i'm proud! #mailboxpride #liberalisme
1	@user @user you don't have the balls to hashtag me as a but you say i am to weasel away.. lumpy tony.. dipshit.
1	makes you ask yourself, who am i? then am i anybody? until ....god . oh thank you god!
0	hear one of my new songs! don't go - katie ellie #youtube #original #music #song #relationship #songwriter
0	@user you can try to 'tail' us to stop, 'butt' we're just having too good of a time! #goldenretriever #animals
0	i've just posted a new blog: #secondlife #lonely #neko

Fig. 2 Head of the sentiment analysis dataset (training)

learning and deep learning models are built and executed to deliver the sentiment analysis using the data that incorporates a broad range of tweets. The block diagram of the recommended model is as given in Fig. 1. Before moving on to developing the analyzer, we first need to perform data cleaning by implementing the following steps. We perform tokenization, remove the unwanted patterns, remove the stop words, and perform stemming. A crucial measure in developing a classifier is determining the features of the input that are pertinent. Then proceed to understand how to encode those features. We extract feature vectors with the help of the Bag-of-words method. Once the data is ready, we build our machine learning model for sentiment analysis and emotion detection. These machine learning models predict sentiment or emotion. We use accuracy, F1 score, and confusion matrix throughout to assess our model's performance.

Introduction to the dataset

The dataset to train our ML model for binary sentiment analysis has 100042 tweets [38]. The dataset which we utilize possesses three columns: 'id', 'sentiment label', and 'sentiment text'. The sentiment label can either be 0 for positive or 1 for negative. In the training dataset, we have three columns present. First is 'id' which is linked to the tweets in the given dataset. The next indicates the tweets collected from diverse



(a)

	Emotion	Text
0	neutral	There are tons of other paintings that I thin...
1	sadness	Yet the dog had grown old and less capable , a...
2	fear	When I get into the tube or the train without ...
3	fear	This last may be a source of considerable disq...
4	anger	She disliked the intimacy he showed towards so...
5	sadness	When my family heard that my Mother's cousin w...
6	joy	Finding out I am chosen to collect norms for C...
7	anger	A spokesperson said : ` Glen is furious that t...
8	neutral	Yes .
9	sadness	When I see people with burns I feel sad, actua...

(b)

Fig. 3 Emotion classification dataset (**a**) Distribution (**b**) Head of the dataset

sources where they indicate the tweet's polarity as positive or negative. The last is a tweet label. The first 15 tweets and labels are shown in Fig. 2. It can be observed that the tweets having words love, proud, new songs, blog are labeled as 0 and the tweets with words offended, lumpy are labeled as 1.

The dataset used to train the model for emotion classification has 7934 tweets [39]. This dataset has 3 columns namely 'id', 'emotion', and 'text'. The emotions are as follows- joy, sadness, neutral, anger, and fear. Figure 3(a) shows the number of data entries in every class. Joy has the maximum number of data entries which are 2326 entries. The column details and some of the tweets with labels are shown in Fig. 3(b).

Preprocessing of the dataset

In data pre-processing, the aim is to perform data cleaning, data integration, data reduction, and data transformation. We start with removing unwanted patterns followed by removing the stop words and performing stemming. Before eliminating stop words, we need to perform tokenization as well. Stop words are words that commonly occur in any natural language. To analyze the textual data and construct natural language processing models we need to remove stop words. Stop words don't add much significance to the meaning of the document. Words like "is", "a", "on", and "the" add no meaning to the statement while parsing it so these stop words. Now after this stemming is performed. Stemming plays a pivotal role in the pipelining course in Natural language processing. The input to the stemmer always needs to be tokenized words. This paper takes the aid of the Bag-of-Words method for feature extraction. It is a technique used to extract features from textual documents. The features can be further utilized for training various ML techniques. It creates a vocabulary of all the distinctive words present in all the documents in the training set. After this, the first task is to split the dataset into training and validation set so that the training and testing of our model can begin before applying it to predict unseen and unlabeled test data.

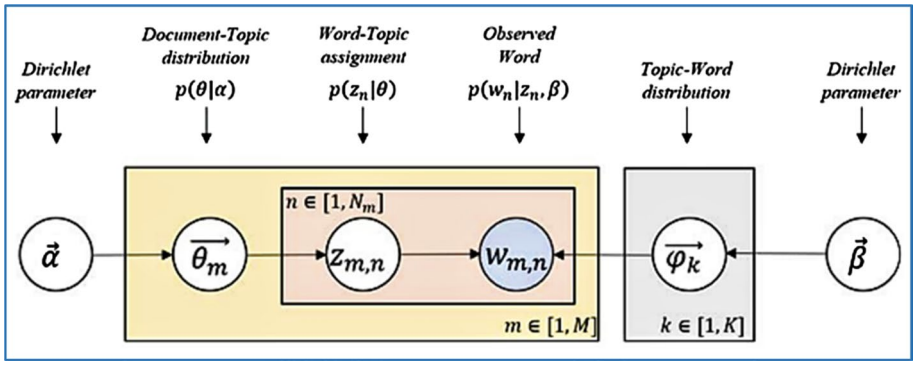


Fig. 4 LDA Architecture

Topic modelling with LDA

The methodology in LDA first constitutes data pre-processing as shown in Fig. 4. A dictionary is created containing the number of times a word appears in the training set and all the anomalies are filtered out. For every document, a dictionary is created reporting how many words and how many times those words appear. LDA has three important hyperparameters. The first one is 'alpha' which outlines a document-topic density factor. The second one is 'beta' which denotes word density in a topic. The third one is 'k', or the number of components signifying the number of topics the document is to be clustered or divided.

Binary sentiment classification

Supervised learning problems can branch into two categories which are regression and classification problems. The problem which the paper addresses come under the classification category because we must classify our results into either the Positive or Negative class. Three models are implemented which are Logistic Regression, Decision Trees, and Random Forest. Pseudocodes of these algorithms are shown in Algorithm 1, 2, and 3. Their performance is compared, and the best possible model is chosen. We used accuracy, F1 score, and confusion matrix throughout to assess our model's performance. Random Forest has the best accuracy and does well in all the other parameters as well when in comparison to the other models.

Algorithm 1 Logistic regression

Precondition: A training set S := $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, features F.

1. **function** Logistic Regression
2. Until convergence
3. $h = H(X,\theta)$// Predicting values from current theta values using logit function
4. gradient $\nabla = \frac{1}{m} mX^T(h - y)$
5. $\theta = \theta - \alpha\nabla$ // update the parameters θ=theta−α∇θ=theta−α∇
6. Compute loss function $J(\theta)$
7. **end function**

Algorithm 2 Decision Tree

Precondition: A training set S:= (x_1, y_1), (x_2, y_2) . . . ,(x_n, y_n), features F, and the number of trees in forest B.

1. **function** Decision tree
2. Calculate the Information gain and Entropy for each attribute.
3. Pick the attribute with the highest information gain, and make it the decision root node.
4. Calculate the information gain for the remaining attributes.
5. Create recurring child nodes by starting splitting at the decision node (i.e for various values of the decision node, create separate child nodes).
6. Repeat this process until all the attributes are covered.
7. Prune the Tree to prevent overfitting.
8. **end function**

Algorithm 3 Random Forest

Precondition: A training set S:= (x_1, y_1), (x_2, y_2) . . . ,(x_n, y_n), features F, and the number of trees in forest B.

1. **function** Random Forest(S, F)
2. H ← ∅ {Assembly of trees, reset to NULL}
3. for i ∈ 1, . . . , B
4. do
5. S(i) ← A bootstrap trial with S
6. hi ← RandomizedTreeLearn(S(i), F)
7. H ← H ∪ {hi}
8. end for
9. return H
10. **end function**
11. **function** RandomizedTreeLearn (S , F)
12. while not done
13. At every node:
14. f ← f U Fragmented on finest feature in F
15. F← F-a
16. **return** the updated tree f
17. end function

Emotion analysis with BERT model

After loading the BERT Classifier and Tokenizer along with the Input modules. The configuration of the loaded BERT model and the fine-tuning to make it ready to make further predictions begins. In this paper, the BERT model has been trained using ktrain to recognize the emotion on text. Text classification is performed with the help of the ktrain library. As shown in Fig. 5, BERT utilizes the features of a Transformer, a capable structure that studies contextual relations in a text with regards to words. In its plain arrangement, a transformer comprises two distinct mechanisms. The first mechanism is an encoder that peruses the input. The second mechanism is a decoder that induces a prediction for the respective assignment. In contrast to directional models, which peruse the input successively, the whole arrangement of words

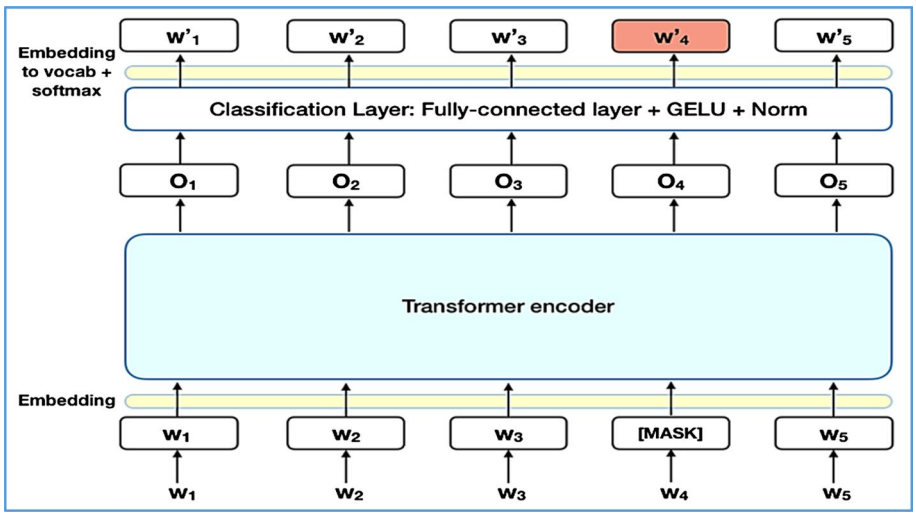


Fig. 5 BERT Architecture

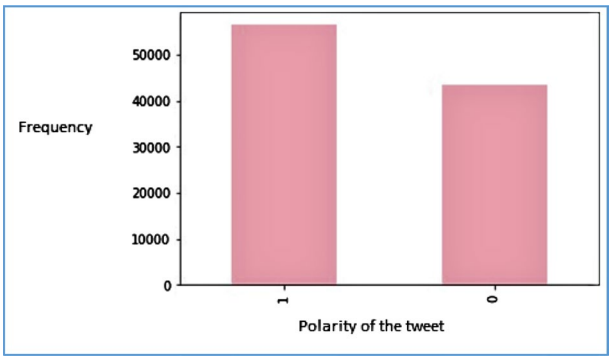


Fig. 6 Number of Negative and Positive Tweets in the training dataset

is delivered at once by the Transformer encoder. Hence, it is regarded as a bidirectional model. However, it is more precise to state it non-directional.

Results and discussion

In this section, we present exploratory analysis, results of topic modeling, binary sentiment analysis using ML algorithms, and emotion detection using the BERT model.

Data exploratory analysis

Figure 6 shows an exploratory analysis of the tweets. Figure 6 depicts the positive and negative tweets in the training dataset. Over here '0' denotes positive tweet and '1' denotes negative tweet. We can observe there are more than 50000 positive tweets and around 40000 negative tweets in the sentiment analysis dataset. In Fig. 7 we are checking the distribution of tweets in the training and testing dataset. The training dataset is shown in pink color whereas the testing dataset is shown in orange color. This graph denotes that there are more tweets in the training dataset and the length is between 0 and 200 characters for both datasets (Fig. 7).

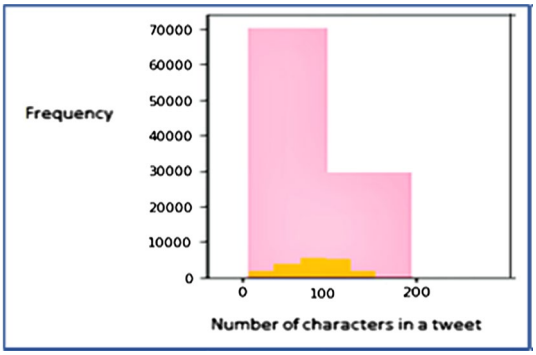


Fig. 7 Distribution of tweets in the training and testing dataset

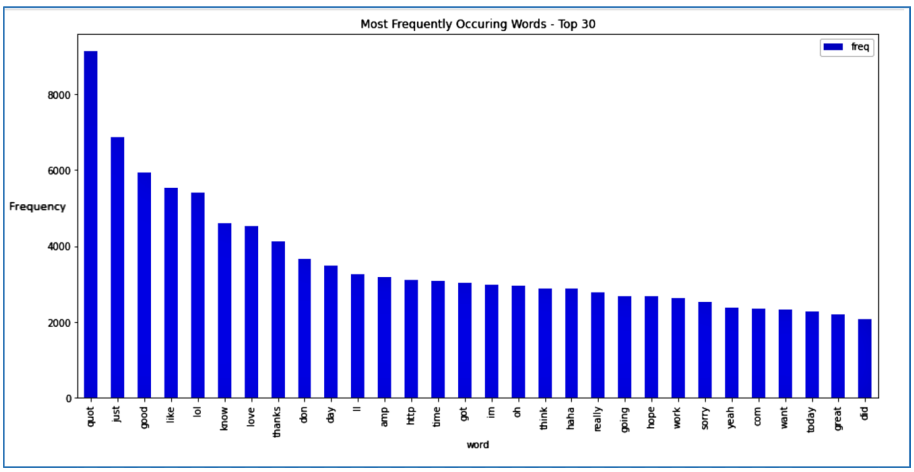


Fig. 8 Top 30 most frequently occurring words

In the bar plot shown in Fig. 8, we can observe the thirty most frequently occurring words. We perform this with the help of the CountVectorizer function. We can observe that the word quot occurs more than 8000 times in the dataset. The word quot is followed by just, good, and like respectively. Word Cloud is the kind of visualization where the most recurrent words are showcased in bigger sizes and the less recurrent words are showcased in relatively small sizes. In Python, we have a package for producing WordCloud. In this paper, we have showcased the top 30 most recurring words in my dataset with the help of WordCloud and Bar plots. The WordCloud in Fig. 9 shows the 30 most frequently occurring words. In WordCloud the word occurring the most commonly appears the largest. Since quot is most recurring it is shown to be the biggest here.

Topic modelling LDA

Topic modeling results are shown in Fig. 10. It depicts the top 10 models and the cluster of words falling each of the topics. We can observe that 0, 1 and 2 are related the college life and some words depict the sad status. Topic modeling is beneficial, but it's tough to comprehend it just by having a glance at the combination of words and statistics. One of the most efficacious ways to interpret data is done with the assistance

Fig. 9 WordCloud (**a**) top 30 words (**b**) most recurring positive words

of visualization. We used PyLDAvis collaborative LDA visualization python package to visualize topic modeling results. PyLDAvis permits us to comprehend the subjects in a topic model. With the assistance of this package, we get to realize the most recurring words in every individual topic along with their occurrence. Moreover, it even demonstrates how related are the topic to each other.

Each bubble in PyLDAvis indicates a topic. The bigger the bubble, the higher fraction of the number of tweets in the corpus is concerning that topic. Blue bars signify the general frequency of each word in the corpus. If no topic is chosen, the blue bars of the most used words will be shown. Red bars give the projected number of times a given term was produced by a given topic. We can conclude from the graph shown in Fig. 11, that the words sad, school, and food are the most recurring words in the dataset. Visualization for topics 1, 3, and 5 can be seen in Figures 12, 13, and 14. This is a very good interpretation to understand the overall orientation of the tweets and sentiments. It is a way to explore the tweets. Further, these topics can also be used to group and label the group of words. With topic modeling analysis of tweets, we could interpret that most of the tweets are belonging to student life with neutral and sad emotions.

Binary sentiment analysis

The sample tweets from the binary sentiment dataset are labeled as '0' or '1' where 0 stands for positive sentiment and 1 stands for negative sentiment as shown in Figures 15 and 16. For example, the tweets having words like "love", "thank", "welcome" etc. have the label '0' and tweets with words "insecure", "lumpy", "devasted" etc. are labeled as '1'. These tweets are pre-processed and are given to ML models for training. Figure 17 depicts the evaluation metrics for trained models built using logistic regression, decision tree, and random forest. The ML algorithms are implemented in 10 randomized experimental runs. We concluded that Random Forest Classifier is a better model than Logistic Regression and Decision Trees due to a high accuracy which was 97.78% on the test dataset. Hence, it is used in the final framework. Predictions obtained for training data using random Forest are shown in Fig. 18. The tweets which are marked as negative need further analysis for stress conditions. In the initial and predicted labels, most of the neutral tweets are classified as positive, hence it will not affect identifying the stress conditions. To confirm this, we also implemented the Vader algorithm which classifies tweets

	topic_0	topic_1	topic_2	topic_3
0	0.108*"bts_meal"	0.148*"food"	0.046*"college"	0.127*"unhappy"
1	0.099*"mcdonaldsindia_decide_spicy_try"	0.064*"school"	0.044*"student"	0.071*"people"
2	0.071*"amp"	0.036*"state"	0.041*"sad_love"	0.041*"sponsor_adidas_unhappy_club"
3	0.044*"assignment_quickly"	0.035*"day"	0.041*"haram_felt_excite_embarrass"	0.041*"fall_shirt_sale"
4	0.044*"preshdeyforyou_fuck"	0.033*"unhappy"	0.041*"sevendless_youngjae_clear_stable"	0.041*"mailsport_man_united_large"
5	0.044*"school_finish"	0.029*"feel"	0.041*"vocal_already_give_acting"	0.040*"school"
6	0.044*"cause_afraid"	0.029*"always"	0.034*"everything"	0.039*"today"
7	0.044*"failure"	0.028*"never"	0.032*"really"	0.028*"good"
8	0.041*"school"	0.028*"two"	0.030*"last"	0.028*"wan_na"
9	0.031*"morning"	0.028*"start"	0.027*"demolarewaju_sad"	0.022*"twitter"
10	0.030*"stop_life"	0.027*"life"	0.027*"evil_people"	0.022*"week"
11	0.030*"friend"	0.027*"high_school"	0.027*"app_everywhere"	0.021*"imagine"
12	0.030*"college_amazing"	0.027*"money"	0.027*"really_careful"	0.018*"student"
13	0.030*"read_tell"	0.024*"meal"	0.027*"umoren_kill"	0.018*"nothing"
14	0.028*"best"	0.021*"save"	0.027*"hear_ini"	0.014*"look"
15	0.022*"hear"	0.021*"cause"	0.027*"amp"	0.014*"hand"
16	0.022*"name"	0.021*"college_student"	0.022*"hour"	0.014*"die"
17	0.021*"month"	0.019*"mean"	0.021*"mind"	0.014*"literally"
18	0.021*"send"	0.014*"send"	0.021*"time"	0.014*"change"
19	0.021*"money"	0.014*"back"	0.021*"still"	0.014*"ago"

(a)

topic_4	topic_5	topic_6	topic_7
0.312*"sad"	0.125*"medium_behave_godi_medium"	0.054*"food"	0.171*"meal"
0.072*"college"	0.125*"yashwantsinha_modi_think_foreign"	0.040*"details"	0.089*"mcdonalds_decide_spicy_bts"
0.064*"really_sad"	0.125*"criticism"	0.040*"upadhyay_college_isolation_center"	0.060*"day"
0.032*"humoren_dead_wicked"	0.125*"india_first_time_face"	0.040*"oxygen_support_functional_verify"	0.059*"new"
0.031*"impend_asteroid_strike"	0.054*"soothe_soul_provide"	0.040*"haramiparindey_delhi_dayal"	0.049*"forkeyus_ever_think"
0.031*"excellion_breaking_dinosaurs_unhappy"	0.054*"rice_added_goodness_walnut"	0.032*"hard_earn"	0.049*"meal_keyu_busy_day"
0.024*"school"	0.054*"cawalnutsindia_comfort_bowl_dal"	0.032*"right_tell"	0.049*"till_night_周柯宇_奥斯卡"
0.022*"local"	0.049*"food"	0.032*"anyone_spend"	0.045*"school"
0.022*"next"	0.036*"eat"	0.032*"money_much"	0.031*"teacher"
0.021*"time"	0.025*"fuck"	0.032*"sacrifice_skip"	0.029*"peace"
0.020*"amp"	0.023*"kill"	0.028*"college"	0.028*"work"
0.017*"love"	0.019*"special"	0.027*"week"	0.025*"quality_food"
0.016*"give"	0.017*"great"	0.025*"amp"	0.025*"sad"
0.016*"head"	0.016*"weird"	0.025*"think"	0.023*"government"
0.011*"year"	0.016*"night"	0.025*"care"	0.022*"post"
0.011*"fixthecountry"	0.009*"day"	0.024*"give"	0.018*"opportunity_thousand"
0.011*"black_tomorrow"	0.009*"little"	0.024*"via"	0.014*"suggest"
0.011*"price"	0.009*"hear"	0.024*"man"	0.014*"tell"
0.011*"spend"	0.009*"name"	0.018*"new"	0.014*"let"
0.010*"military"	0.009*"right"	0.017*"year"	0.012*"special"

(b)

Fig. 10 Topic Modelling LDA results (**a**) topic 1 to topic 4 (**b**) topic 5 to topic 8 (**c**) topic 9 to topic 10

into positive, negative, and neutral using a sentiment dictionary. The output of the Vader sentiment analysis is shown in Fig. 19. It can be observed that compound values are from −1 to 1 and values between −0.5 to 0.5 are identified as neutral. We also tested the random forest model for random tweets and one of the outputs where the tweet is classified as positive is shown in Fig. 20.

BERT model results

Since the emotion classifier has 5 classes namely- joy, sad, neutral, angry, and fear. It will be categorizing the tweets in those emotions only. The training of the model and accuracies obtained at different epochs are shown in Fig. 21. It can be observed that the training accuracy of the BERT model after using one cycle policy at a learning rate of 0.00002 is 94%. The evaluation of the model on the test set of 6000 tweets is shown in Fig. 22. The figure depicts micro evaluation metrics: accuracy, F1 score,

topic_8	topic_9
0.051*"help"	0.196*"school"
0.044*"gold_belonging_venezuela_value"	0.043*"always"
0.044*"support_illegally_retain_ton"	0.036*"extremely_sad"
0.037*"ceder_international"	0.036*"life_wrong"
0.037*"julianaahua_missing"	0.036*"note_feel"
0.037*"nimota_pascaline_chocolate_seen"	0.036*"theplacardguy_serious"
0.030*"someone"	0.036*"people_lay"
0.030*"school"	0.032*"thank"
0.030*"abroad_wait"	0.032*"year"
0.030*"surgery_doctor"	0.030*"today"
0.030*"line_receive"	0.027*"food"
0.030*"minsugahq_people"	0.027*"free"
0.030*"joke_afraid"	0.027*"taemin"
0.030*"armys"	0.025*"happy_unhappy"
0.024*"meal"	0.023*"poor_people"
0.023*"mcdonaldscanada_decide_spicy"	0.018*"person"
0.023*"remember"	0.018*"dinner"
0.023*"bts_meal"	0.018*"bad"
0.022*"thought"	0.017*"every"
0.022*"harm_anyone"	0.017*"meal_deal"

(c)

Fig. 10 continued

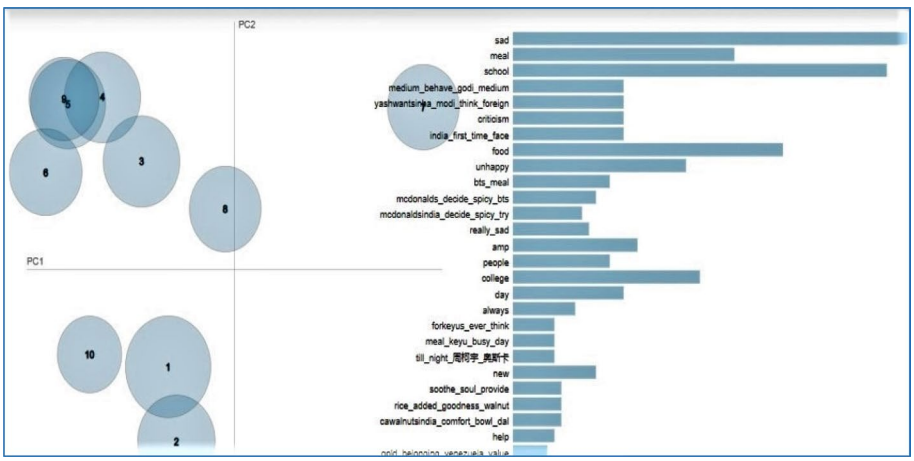


Fig. 11 Representation of the top salient terms in the dataset

precision, recall, and sensitivity for each class. It can be observed that the model has a good F1-score and accuracy for all the classes. The model has a macro average accuracy of 94%, and a macro F1-score of 83%. It indicates that the model is not overfitting.

The web portal is designed for topic modeling and emotion detection. The model is used to classify emotions in the web portal. Figures 23, 24, 25, 26, 27 indicate the classification of the emotion for the post given as input. These depict the prediction after entering the text. The figures depict the output for each of the emotions for corresponding input texts entered as input. When the input is classified as negative. Sadness, anger, or fear, we further analyze more tweets to classify the user as stres sed.

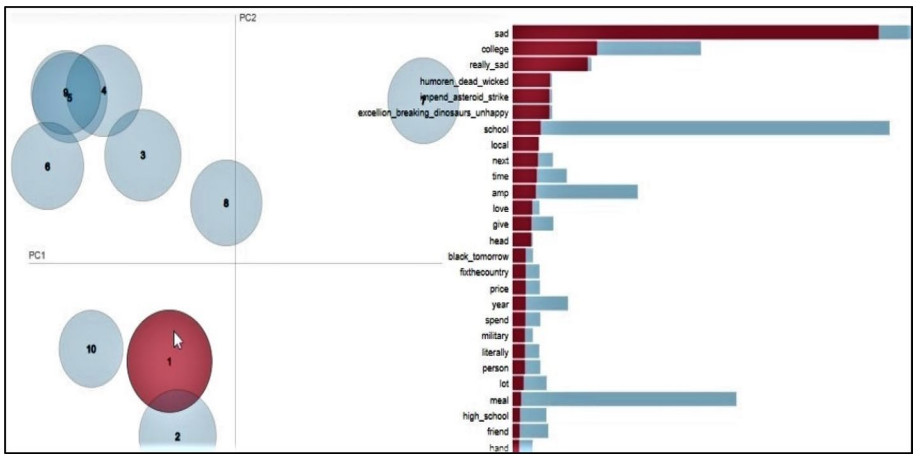


Fig. 12 Topic 1's most frequent words (marked in red)

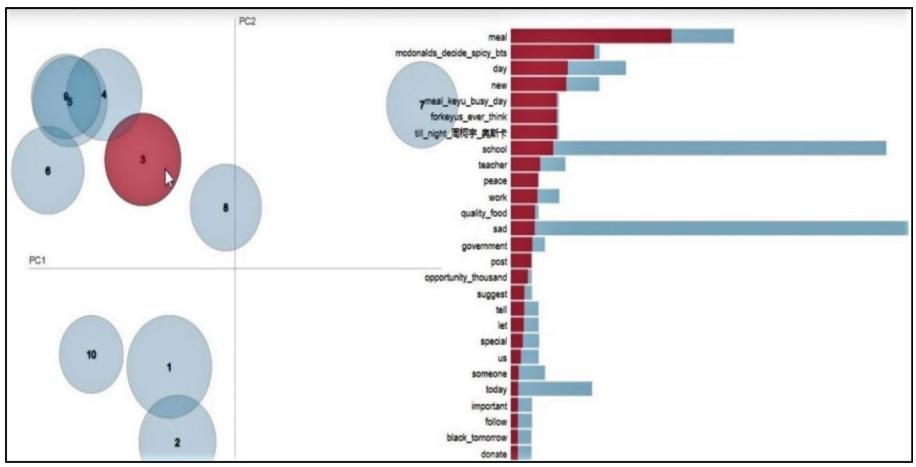


Fig. 13 Topic 3's most frequent words (marked in red)

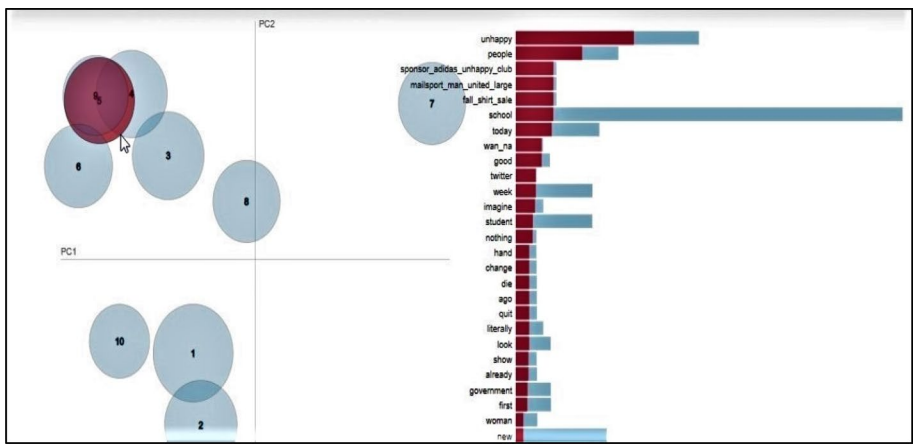


Fig. 14 Topic 5's most relevant words

id	label	cleaned_tweet
1	0	user father dysfunct selfish drag kid dysfunct run
2	0	user user thank lift credit use caus offer wheelchair van pdx disapoint get thank
3	0	birthday majesty
4	0	model love u take u time ur
5	0	facts guide societiy motive
6	0	huge fan fare big talk leav chao pay dispute get allshowandnogo
7	0	user camp tomorrow user user user user user user danni
8	0	next school year year exam think school exam hate imagine actorslif revolutionschool girl
9	0	love land allin cave champion cleverland clevelandcavali
10	0	user user welcome gr
11	0	ireland consume price index mom climb previous may blog silver gold forex
12	0	selfish orlando standwithorlando pulseshoot orlandoshoot biggerproblem selfish heabreak valu love
13	0	get see daddy today day gettingf
16	0	ouch junior angri got junior yugyoem omg
17	0	thank paner thank positive
19	0	friday smile around via ig user user cooki make people
20	0	know essential oil made chemical

Fig. 15 Tweets with label 0 in the training set

id	label	cleaned_tweet
14	1	user can call michigan middle school build wall chant tcot
15	1	comment australia opkillingbay seashepherd helpcovedolphin thecov helpcovedolphin
18	1	retweet agree
24	1	user user lumpy say prove lumpy
35	1	unbelieve st century need something like neverump xenophobia
57	1	user let fight love peace
69	1	white establish blk folx run around love promote great
78	1	user hey white people call peopl white user race ident med
83	1	altright use amp insecur lure men whitesupremaci
112	1	user interest linguistic address race amp racism power raciolinguist bring
115	1	user user mock obama black user user user user brexit
132	1	peopl protest trump republican trump fuher amp
152	1	ye call michelleobama gorilla racist long thought black peopl bet
157	1	smaller hand show barri probabl lie knick game suck golf
168	1	user user point one finger user million point right back jewishsupremacist
193	1	might libtard libtard sjw liber polit
211	1	user take trash america vote hate vote vote vot
233	1	hold open door woman woman nice thing even tri deni
264	1	user man ran governor ny state biggest african american population

Fig. 16 Tweets with label '1' in the training set

Conclusion

Sentiment and emotion analysis is an area of learning to examine opinions expressed in the text on numerous social media websites. In this paper, we presented the exploratory analysis of user tweets using LDA topic modeling and visualization. We emphasized the importance of data visualizations as it helps us in getting an apt understanding of our data. In this research work, extracted tweets are analyzed by using LDA to settle on the number of topics and the percentage of a word in a specific topic. The outcomes presented that the extracted topics display a significant structure in the data. We applied and evaluated machine learning algorithms for sentiment analysis, the BERT model for emotion analysis. Models are fine-tuned for the sentiment classification task of 5 different classes—Joy, Sadness, Neutral, Anger, and Fear. We have verified the classification competence in NLP supported

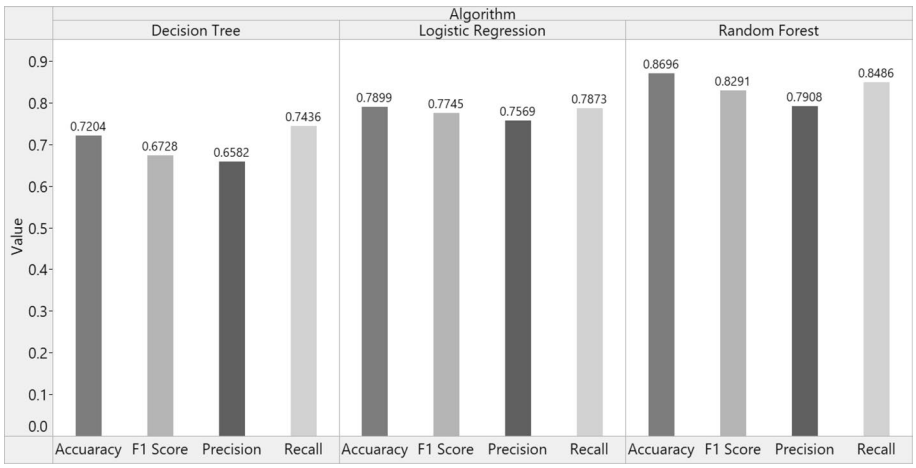


Fig. 17 Evaluation metric for binary classification on training data for ML models

Column1	id	cleaned_tweet	Predicted
0	31963	studiolife aislife require passion dedication willpower find newmaterial	0
1	31964	user white supremacist want everyone see new bird movie why	0
2	31965	safe way heal acn altwaystoh healthi heal	0
3	31966	hp curs child book reserve already yes no harrypott pottermor favoritee	0
4	31967	birthday amaze hilarious nephew eli ahmir uncl dave love miss	0
5	31968	choos momtip	1
6	31969	something inside die eye ness smokeyey tire lone sof grung	0
7	31970	finish tattoo ink ink loveit thank aleee	0
8	31971	user user user never understand dad left young deep in the feel	1
9	31972	delicious food lovelif capetown mannaepicur restur	0
10	31973	dayswast narcosi infinite ep make aware grind neuro bass lifestyle	1
11	31974	one world greatest spo event leman teamaudi	1
12	31975	half way website allgoingwell	0
13	31976	good food good life enjoy call garlic bread iloveit	0
14	31977	stand behind guncontrolpleas senselessshoot takethegun comicrelief stillsad	1
15	31978	ate ate ate jamaisasthi fish curri prawn hilsa foodfestiv food	0
16	31979	user got user limit edit rain shine set today user user user user	0
17	31980	love amp hug amp kiss keep babi parent healthcar	0
18	31981	girl sun fave london unit kingdom	0
19	31982	thought factory bbc neutral right wing fascism polit media blm brexit trump leadership gt	1
20	31983	hey guy tommorow last day exam happy yay	1
21	31984	user user user levyrroni recuerdo memori recuerdo friend life triunfodelamor	1

Fig. 18 Output Excel file after Binary Sentiment Analysis using Random Forest

by deep contextual language models like BERT with an accuracy of 94%. The paper aimed to employ these techniques for detecting users' mental stress from twitter's social media facts and figures. After implementing 3 machine learning techniques in the overall of 10 randomized experimental runs, we concluded that Random Forest Classifier is a better algorithm than Logistic Regression and Decision Trees with an accuracy of 97.78%. We developed a web portal that takes the text posted by the user as the input and identifies the emotion. The portal has shown accurate classifications for any given tweet. The work can be further extended by combining other health parameters to monitor mental health.

Sino	tweet	compound	neg	neu	pos
0	studiolife aislife require passion dedication willpower find newmaterial	0.4588	0	0.7	0.3
1	user white supremacist want everyone see new bird movie why	0.2023	0	0.714	0.286
2	safe way heal acn altwaystoh healthi heal	0.4404	0	0.674	0.326
3	hp curs child book reserve already yes no harrypott pottermor favoritee	0	0	1	0
4	birthday amaze hilarious nephew eli ahmir uncl dave love miss	0.5574	0.108	0.608	0.284
5	choos momtip	0	0	1	0
6	something inside die eye ness smokeyey tire lone sof grung	-0.7184	0.429	0.571	0
7	finish tattoo ink ink loveit thank aleee	0.3612	0	0.706	0.294
8	user user user never understand dad left young deep in the feel	0	0	1	0
9	delicious food lovelif capetown mannaepicur restur	0	0	1	0
10	dayswast narcosi infinite ep make aware grind neuro bass lifestyle	0	0	1	0
11	one world greatest spo event leman teamaudi	0.6369	0	0.588	0.412
12	half way website allgoingwell	0	0	1	0
13	good food good life enjoy call garlic bread iloveit	0.8402	0	0.4	0.6
14	stand behind guncontrolpleas senselessshoot takethegun comicrelief stillsa	0	0	1	0
15	ate ate ate jamaisasthi fish curri prawn hilsa foodfestiv food	0	0	1	0
16	user got user limit edit rain shine set today user user user user	0	0	1	0
17	love amp hug amp kiss keep babi parent healthcar	0.8779	0	0.409	0.591
18	girl sun fave london unit kingdom	0.4404	0	0.633	0.367
19	thought factory bbc neutral right wing fascism polit media blm brexit trump	0.2732	0	0.861	0.139
20	hey guy tommorow last day exam happy yay	0.5267	0	0.673	0.327
21	user user user levyrroni recuerdo memori recuerdo friend life triunfodelamo	0.4939	0	0.738	0.262

Fig. 19 Output Excel file after Binary Sentiment Analysis using Vader Algorithm

```
good food, good life , #enjoy
(1, 4)
Result: Positive
```

Fig. 20 Sentiment Analysis of a tweet

```
learner.fit_onecycle(2e-5, 3)


begin training using onecycle policy with max lr of 2e-05...
Epoch 1/3
1323/1323 [==============================] - 1054s 781ms/step - loss: 1.2734 - accuracy: 0.4646 - val_loss: 0.5644 - val_accura
cy: 0.7987
Epoch 2/3
1323/1323 [==============================] - 1037s 784ms/step - loss: 0.4590 - accuracy: 0.8517 - val_loss: 0.5106 - val_accura
cy: 0.8170
Epoch 3/3
 935/1323 [====================>.........] - ETA: 4:35 - loss: 0.1962 - accuracy: 0.9422
```

Fig. 21 Bert model created with an accuracy of 94 percent for overall classification (Macro)

Future work

Effectual analysis of policy opinionated content

The future scope of the paper is to develop a system that not only detects stress but also analyses the topic of discussion in a particular tweet. This could work as a survey system. It would provide a better solution on every debatable topic and tell the popular choice/verdict in areas like politics and news. This will help us efficiently analyze stress and express opinions for prevailing social issues.

Detection of spam and non-spam tweets

This paper could help analyze if a tweet is spam or non-spam. This could potentially help naïve Twitter users be aware of spam accounts which could be harmful to a lot

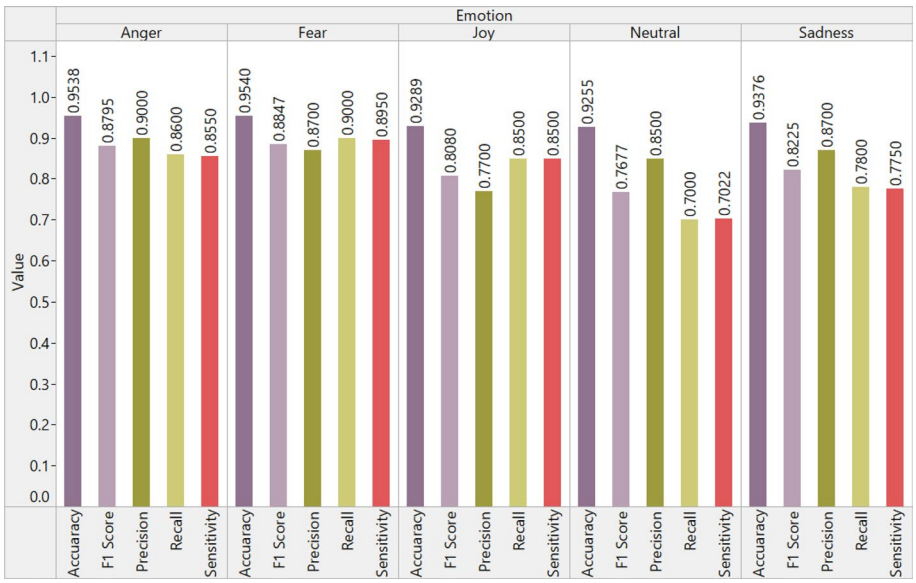


Fig. 22 Performance evaluation of test set using BERT model (Micro level)

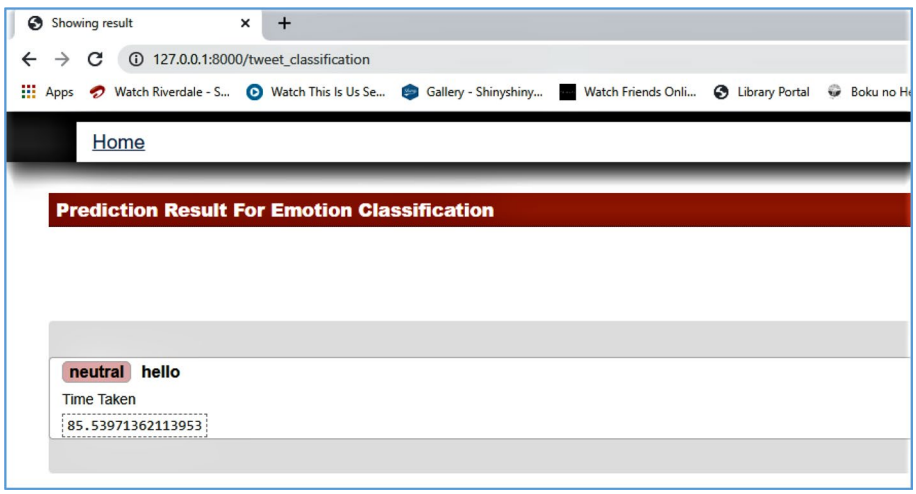


Fig. 23 Neutral Tweet with the time taken 85 s

of Twitter users. The non–spam tweets can also be further classified to make sure the ones which are damaging are removed from the Twitter platform.

Improving sentiment word identification algorithm

With social media, there are a lot of impediments. A tweet can have abbreviations, slang, and jargon which is difficult to interpret. This project can be further used to perform analysis on short sentences and abbreviations to get a better idea. Additionally, people should work on the generation of a high content lexicon database. There should also be successful handling of bi-polar sentiments. All these features combined would help develop an astounding analyzing tool.

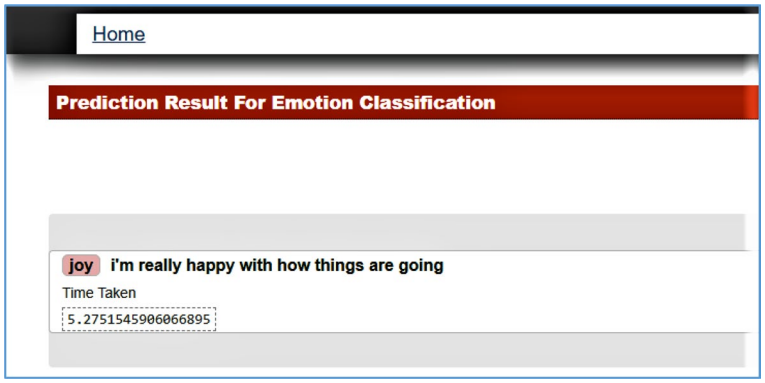


Fig. 24 Joy Tweet with the time taken 5 s

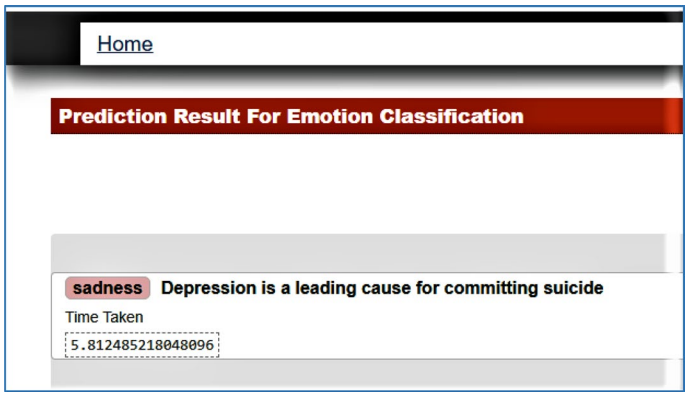


Fig. 25 Sadness Tweet with the time taken 5 s

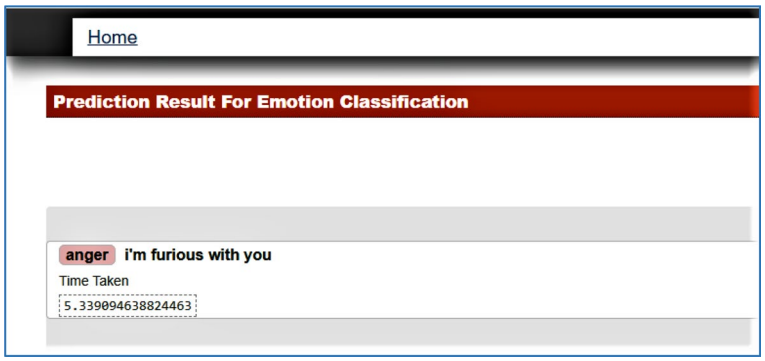


Fig. 26 Anger Tweet with the time taken 5 s

Dynamic topic model

A Dynamic Topic Model will examine the fluctuations of subjects done over time; it is likewise important to consider the addition of time-varying information. Executing a topic modeling outline that will allow the incorporation of supplementary data will

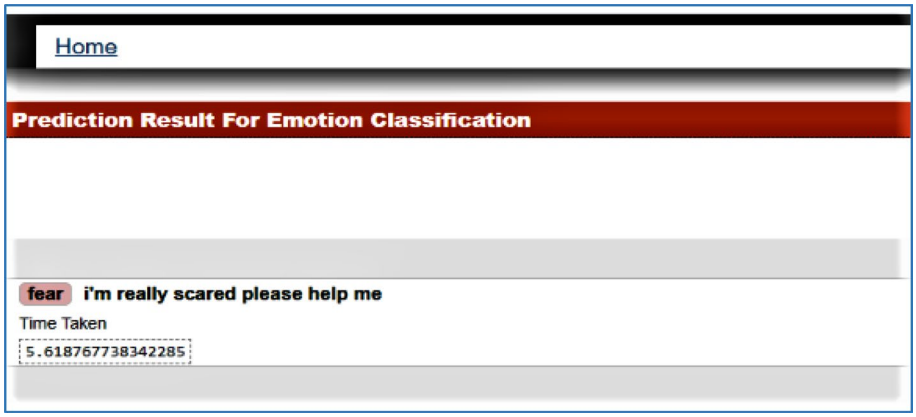


Fig. 27 Fear Tweet with the time taken 5 s

produce an advantageous potential in the turf of publicizing research. Additionally, integrating some method of direction during topic generation can help interpret the derivative solutions.

Acknowledgements
Nil.

Authors' contributions
TN: The author has designed and implemented all the phases of the project. GA: The author has guided throughout the process of the project and has performed result analysis. TA: The author has finalized the structure and content of the manuscript and has done the proofreading. All authors read and approved the final manuscript.

Authors' informations
Tanya Nijhawan is a student in the Department of Electronics and Communication Engineering, Manipal Institute of Technology (MIT), MAHE, Manipal, India. Her research interests include Data Science, Data Mining, Machine Learning
Girija Attigeri is currently Assistant Professor-Selection Grade in the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. She has received B.E. and M. Tech. degrees from the Visvesvaraya Technological University, Karnataka, India. She has 15 years of experience in teaching and research. She has received his Ph.D. from the Manipal Institute of Technology, Karnataka, India. His research interests span big data analytics, Machine Learning, and Data Science. She has around 10 publications in reputed international conferences and journals
Ananthakrishna Thalengala has received his M.Sc. degree in 1998 in Electronic Science from Mangalore University, India, M. Tech. degree in 2004 in Computer Cognition Technology, from Mysore University, India, and a Ph.D. degree in 2019 from Manipal Academy of Higher Education (MAHE), Manipal, India. Since 2004 he is with Manipal Institute of Technology (MIT), MAHE, Manipal, India, where he is currently an Assistant Professor in the Department of Electronics and Communication Engineering. He is a senior member of IEEE, and his areas of interest include Signal processing, Pattern classification, and Machine learning.

Funding
Not applicable.

Availability of data and materials
The sources of the data are cited in the paper. They are 31st and 32nd references.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Author details
[1]Depertment of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India. [2]Depertment of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India.

Received: 2 November 2021 Accepted: 6 February 2022
Published online: 20 March 2022

References

1. Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. Inform Fusion. 2019;52(1):290–307.
2. Liu B, Zhang L. A survey of opinion mining and sentiment analysis. Boston: Springer US. 2012; p. 415–463.
3. Munikar M, Shakya S, Shrestha A. Fine-grained sentiment classification using BERT. Artif Intell Transform Business Society. 2019;2019:1–5. https://doi.org/10.1109/AITB48515.2019.8947435.
4. Wang B, Liu Y, Liu Z, Li M, Qi M. Topic selection in latent Dirichlet allocation, 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). 2014. p. 756–760. https://doi.org/10.1109/FSKD.2014.6980931.
5. Alexander P, Patrick P. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC. 2010.
6. Jianqiang Z, Xiaolin G. Comparison research on text pre-processing methods on Twitter sentiment analysis. IEEE Access. 2017;5:2870–9. https://doi.org/10.1109/ACCESS.2017.2672677.
7. Pradha S, Halgamuge MN, Vinh NQT. Effective text data preprocessing technique for sentiment analysis in social media data, 2019 11th International Conference on Knowledge and Systems Engineering (KSE). 2019. p. 1–8. https://doi.org/10.1109/KSE.2019.8919368.
8. Deepa DR, Tamilarasi A. Sentiment analysis using feature extraction and dictionary-based approaches, 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC). 2019. p. 786–790. https://doi.org/10.1109/I-SMAC47947.2019.9032456.
9. Chaturvedi S, Mishra V, Mishra N. Sentiment analysis using machine learning for business intelligence, 2017 IEEE International Conference on power, control, signals, and instrumentation engineering (ICPCSI). 2017. p. 2162–2166. https://doi.org/10.1109/ICPCSI.2017.8392100.
10. Ho J, Ondusko D, Roy B, Hsu DF. Sentiment analysis on tweets using machine learning and combinatorial fusion, 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). 2019. p. 1066–1071. https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00191.
11. Apoorv A, Boyi X, Ilia V, Owen R, Rebecca P. Sentiment analysis of Twitter Data. Proceedings of the Workshop on Languages in Social Media. 2011.
12. Peddinti MK, Chintalapoodi P. Domain adaptation in sentiment anlaysis of Twitter, in Analyzing Microtext Workshop, AAAI, 2011.
13. Dmitry D, Oren T, Ari R. Enhanced sentiment learning using twitter hashtags and smileys. Coling 2010—23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2. 2010; 241–249.
14. Anupriya P, Karpagavalli S. LDA based topic modeling of journal abstracts. Int Conf Adv Comput Commun Syst. 2015;2015:1–5. https://doi.org/10.1109/ICACCS.2015.7324058.
15. Xiance S, Maosong S. Tag-LDA for scalable real-time tag recommendation. J Comput Inform Syst. 2008;6:23.
16. Krestel R, Fankhauser P. Personalized topic-based tag recommendation. Neurocomputing. 2012;76:61–70. https://doi.org/10.1016/j.neucom.2011.04.034.
17. Peters ME, Neumann M. Deep contextualized word representations. 2018.
18. Radford A, Narasimhan K. Improving language understanding by generative pre-training. 2018.
19. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, vol 1. Minneapolis; 2019. p. 4171–4186. https://doi.org/10.18653/v1/n19-1423.
20. Jin Z, Lai X, Cao J. Multi-label sentiment analysis base on BERT with modified TF-IDF, 2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN), 2020. https://doi.org/10.1109/ISPCE-CN51288.2020.9321861.
21. Zubair M, Aurangzeb K, Shakeel A, Maria Q, Ali KI, Quan Z. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. Plos One. 2017;12(2):e0171649.
22. Zeng D, Dai Y, Li F, Wang J, Sangaiah AK. Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism. J Intell Fuzzy Syst. 2019;36(5):3971–80. https://doi.org/10.3233/JIFS-169958.
23. Alec G, Richa B, Lei H. Twitter sentiment classification using distant supervision. Processing. 2009; 150.
24. Alexandra B, Ralf S, Mijail K, Vanni Z, van der Erik G, Matina H, Bruno P, Jenya B. Sentiment analysis in the news. proceedings of LREC. (n-1). 2013
25. Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, USA, 43–48. [N-2]. 2005.
26. Boiy E, Moens MF. A machine learning approach to sentiment analysis in multilingual web texts. Inf Retrieval. 2009;12:526–58. https://doi.org/10.1007/s10791-008-9070-z[N+1].
27. Li F, Huang M, Zhu X. Sentiment analysis with global topics and local dependency. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI Press. 2010. 1371–1376.

28. Arya V, Mishra AK. Machine learning approaches to mental stress detection: a review. Ann Optimization Theory Pract. 2021;31(4):55–67.
29. Aldarwish MM, Ahmad HF. Predicting Depression Levels Using Social Media Posts, 2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS), Bangkok. 2017. pp. 277–280. https://doi.org/10.1109/ISADS.2017.41.
30. Cho G, Yim J, Choi Y, Ko J, Lee SH. Review of machine learning algorithms for diagnosing mental illness. Psychiatry Invest. 2019;16(4):262–9.
31. Baheti RR, Kinariwala SA. Survey: sentiment stress identification using tensi/strength framework. Int J Sci Res Eng Dev. 2019;2(3):1–8.
32. Deshpande M, Rao V. Depression detection using emotion artificial intelligence, 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India. 2017. p. 858–862
33. Zucco C, Calabrese B, Cannataro M. Sentiment Analysis and Affective Computing for Depression Monitoring. In 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). New York: IEEE. 2017. p. 1988–1995.
34. Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big Data: Deep Learning for Financial sentiment analysis. J Big Data. 2018;5(1):1–25.
35. Bandari S, Bulusu VV. Survey on ontology-based sentiment analysis of customer reviews for products and services. In Data Engineering and Communication Technology. Springer: Singapore. 2020; p. 91–101.
36. Trupthi M, Pabboju S, Gugulotu N. Deep Sentiments Extraction for Consumer Products Using NLP-Based Technique. In Soft Computing and Signal Processing. Springer: Singapore; 2019. p. 191–201.
37. Hammou BA, Lahcen AA, Mouline S. A distributed ensemble of deep convolutional neural networks with random forest for big data sentiment analysis. In International Conference on Mobile, Secure, and Programmable Networking. Springer: Cham; 2019. p. 153–162.
38. Vardhanapu K. Sentiment analysis, IEEE Dataport. 2020. https://doi.org/10.21227/e2aq-xv12.
39. Damian. Detecting Emotions in Text, https://data.world/damof/detecting-emotions-in-text. 2021.

Publisher's Note