Stress Detection using Natural Language Processing and Machine Learning over social Interactions

Tanya Nijhawan[1] Girija Attigeri[2*] Ananthakrishna T[1]
tanyanijhawan74@gmail.com, girija.attigeri@manipal.edu, anantha.kt@manipal.edu
[1]Depertment of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Karnataka, India
[2]Depertment of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Karnataka, India

ABSTRACT

Cyberspace is a vast soapbox for people to post anything that they witness in their day-to-day lives. Subsequently, it can be used as a very effective tool in detecting the stress levels 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 successfully accomplish sentiment analysis with the aid of machine learning algorithms. We take the help of a capable deep learning pre-trained model called BERT to solve the problems which come with sentiment classification. The BERT model outperforms a lot of other well-known models for this job without any sophisticated architecture. We also adopted Latent Dirichlet Allocation which is an unsupervised machine learning method that's skilled in 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 predict which topic is linked to the textual data. With the aid of the models suggested, we will be able to detect the emotion of users online. We are primarily working with Twitter data because Twitter is a website where people express their thoughts often. In conclusion, this proposal is for the well- being of one's mental health. The results are evaluated using various metric at macro and micro level and indicate that the trained model detects the status of emotions bases on social interactions.

Keywords: Decision Tree, Latent Dirichlet Algorithm, Logistic Regression, Machine Learning, Natural Language Processing, Random Forest, Sentiment Analysis, Topic Modelling

Springer

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 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. Social media analytics 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.

To perform the analysis 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. We use the BERT model to do a 5-class emotion classification. The emotions are joy, sadness, neutral, fear, and angry.

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 [1]. 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 machine learning 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
- Develop a Django-based web application that receives inputs from a user and then accordingly generates a prediction.
- Develop a system 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 [4] 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.

Agarwal et al. [5] proposed a 3-way model for categorizing sentiments 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 tree kernel- centered model, tweets were chosen to be represented in the form of a tree. While implementing 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 came to the conclusion 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. [6] 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., [7] 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, [8] 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. [9] 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. [10] 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 [11], OpenAI GPT [12], and BERT [13] 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 base and this can be fine-tuned to respond to the particular NLP task. This process is better than performing the training of the model from the basics. [14] Zubair [15] et al. introduced a technique enhanced by lexicons. It was projected to be centered on a classification scheme which was rule based. 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. [16] On the other hand, pre-trained language models save a lot of time by achieving the same result quickly. They are easy to incorporate and there's not as much labeled data required.

S. No	Researcher	Paper	Technique	Performances
1.	Alexander Pak, Patrick Paroubek	[4]	Naïve Bayes Sentiment classifier with multinomial features	High accuracy, Low decision value
2.	Alec Go, Richa Bhayani, Lei Huang	[21]	Naïve Bayes classifier, Mutual information measure for feature selection	Accuracy: 81%
3	Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov	[22]	WordNet-lexicon	Accuracy: 82% Improvement in the baseline 21%
4.	Jonathon Read	[23]	SVM, Naïve Bayes	Accuracy: 70%
5	Erik Boiy & Marie-Francine Moens	[24]	Integrated approach: ML, Information retrieval, NLP	Accuracy: 83% (English texts),70% accuracy (Dutch texts), 68% (French texts)
6	Fangtao Li, Minlie Huang, Xiaoyan Zhu	[25]	Dependency- Sentiment, LDA, Markov chain	Accuracy: 70.7% with on 10-fold cross-validation test set: 800 reviews

Table 1 Comparison of different approaches in sentiment analysis

RESEARCH DESIGN AND METHODOLOGY

The research makes use of both secondary and primary data sources. It is a cross-sectional study to know the impact of social and emotions associated with the social media data and usefulness of the same. The research is both a quantitative and qualitative study as we aim 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. A hybrid model has been put forward and then

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 [26] [27]. 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. [28]. Once the data is ready, we build our machine learning model for sentiment analysis and emotion detection. [29] [30] These machine learning models predict sentiment or emotion. We use accuracy, F1 score, and confusion matrix throughout to assess our model's performance.

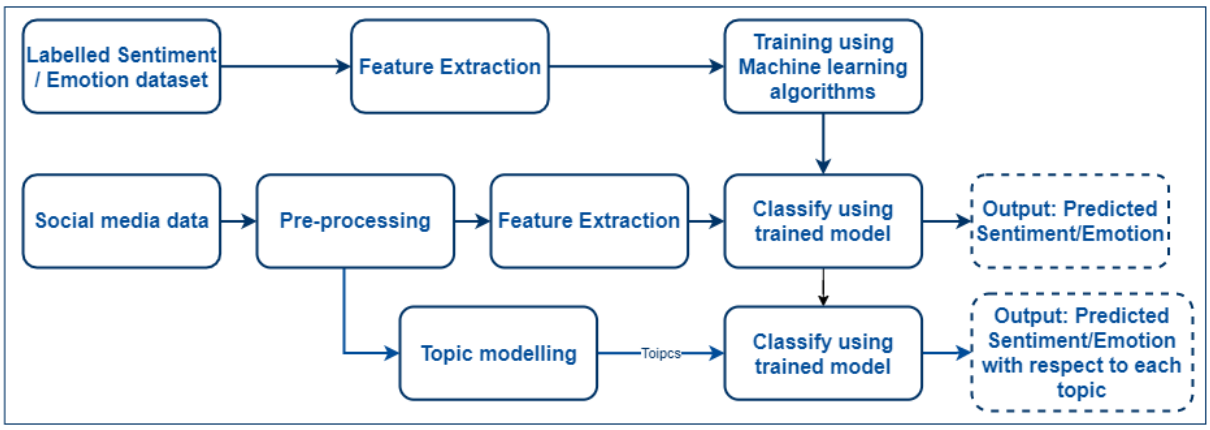

Figure 1 Sentiment Analysis Methodology

Introduction to the Dataset

The dataset to train our machine learning model for binary sentiment analysis has 100042 tweets. The dataset which we utilize possesses three columns: 'id', 'sentiment label', and 'sentiment text'. The sentiment label can either be 0 for negative or 1 for positive. Label '1' characterizes the tweet as positive and label '0' denotes the tweet as 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 sources where they indicate the tweet's polarity as positive or negative. The last is a tweet where label '0' is of negative sentiment while a tweet with label '1' is of positive sentiment. The dataset used to train the model for emotion classification has 7934 tweets. This dataset has 3 columns namely 'id', 'emotion', and 'text'. The emotions are as follows- joy, sadness, neutral, anger, and fear. Figure 6 shows the number of data entries in every class. Joy has the maximum number of data entries which are 2326 entries.

```
[6]: train.head()
```

[6]:

	id	label	tweet
0	1	0	is so sad for my APL frie...
1	2	0	I missed the New Moon trail...
2	3	1	omg its already 7:30 :O
3	4	0	.. Omgaga. Im sooo im gunna CRy. I'...
4	5	0	i think mi bf is cheating on me!!! ...

Figure 2 Head of the sentiment analysis dataset (training)

```
size of training set: 7934
size of validation set: 3393
joy         2326
sadness     2317
anger       2259
neutral     2254
fear        2171
Name: Emotion, dtype: int64
```

(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)

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

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 machine learning techniques. It creates a vocabulary of all the distinctive words present in all of 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.

Topic Modelling with LDA

The methodology in LDA first constitutes data pre-processing. 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.

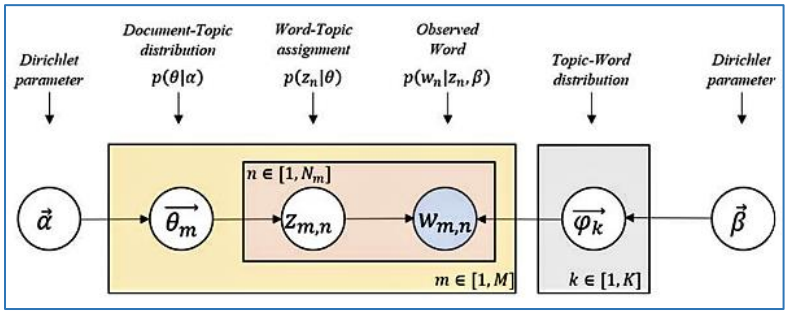

Figure 4 LDA Architecture

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 Algorithm1, 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 := (x1, y1), (x2, y2) . . . ,(xn, yn), features F, and number of trees in forest B.

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 := (x1, y1), (x2, y2) . . . ,(xn, yn), features F, and 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 := (x1, y1), (x2, y2) . . . ,(xn, yn), features F, and 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. 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 of an encoder that peruses the input. The second mechanism is of 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 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.

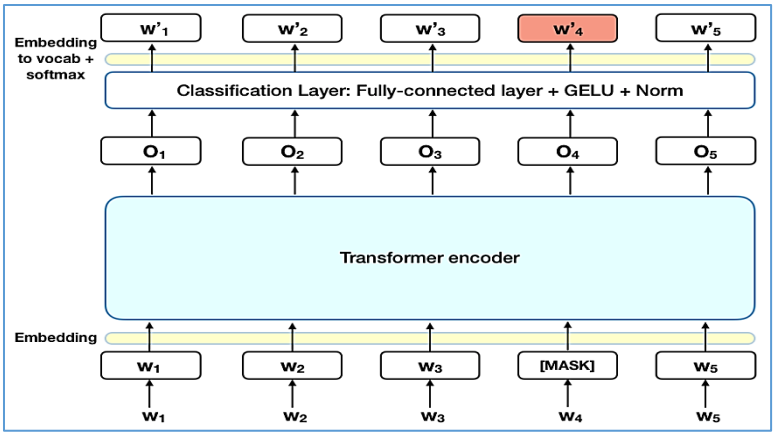

Figure 5: BERT Architecture

RESULTS AND DISCUSSION

In this section, we present the results of each of the phases discussed in the previous section.

Data Exploratory Analysis

In the figure below we are displaying the positive and negative tweets in the training dataset. Over here '1' denotes positive tweet and '0' denotes negative tweet. We can observe there are more than 50000 positive tweets and around 40000 negative tweets in the dataset.

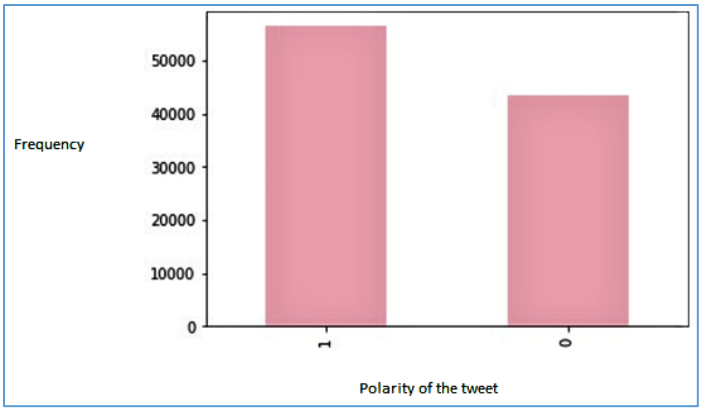

Figure 6: Positive and negative tweets in the training dataset

In figure 10, 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 to 200 characters for both datasets.

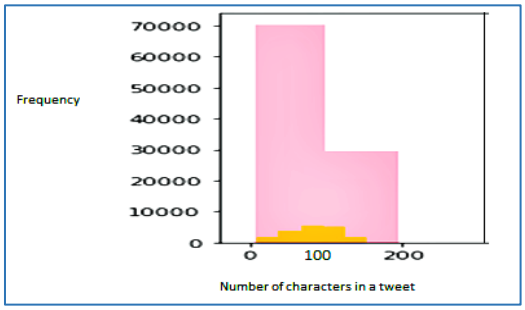

Figure 7 Distribution of tweets in the training and testing dataset

In the bar plot shown in Figure 11, 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 below 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.

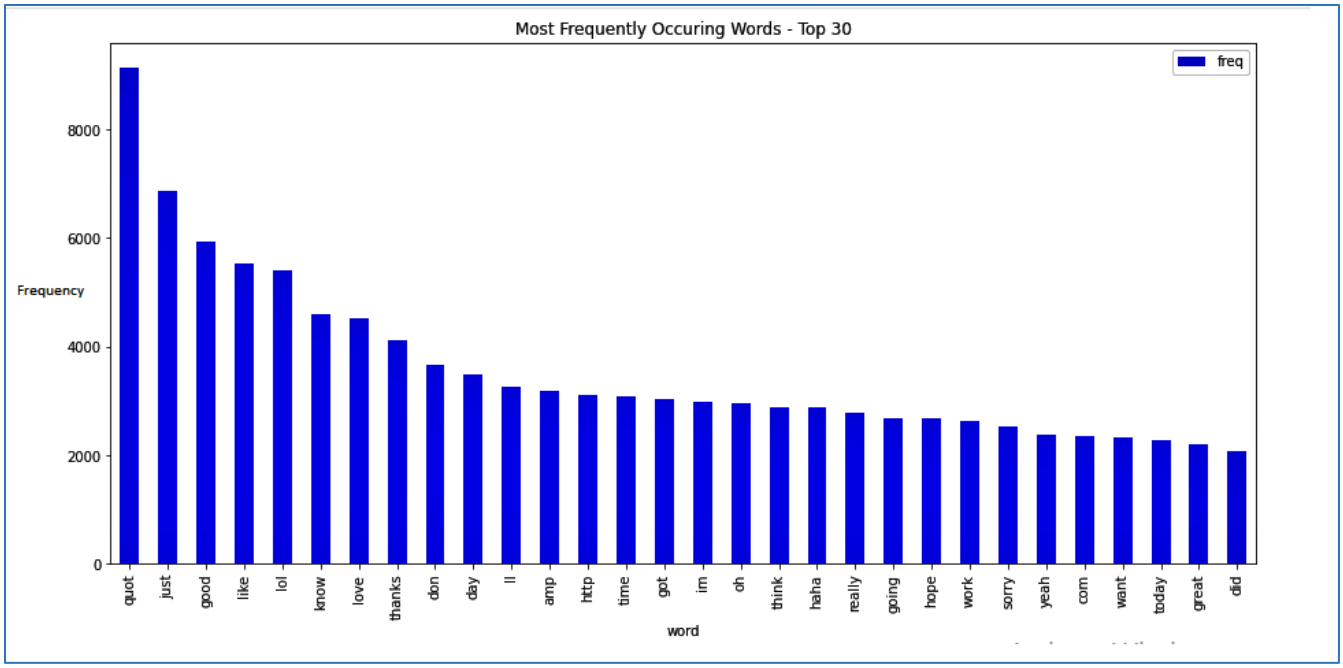

Figure 8 Top 30 most frequently occurring words

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

Topic Modelling LDA

We can enter the path location of the excel file which we want to get analyzed. After entering the path location of the file we'll get the topic modeling results as shown in Figure 20. 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_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)

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)

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

PyLDAvis is a collaborative LDA visualization python package. 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 of visualization. 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.

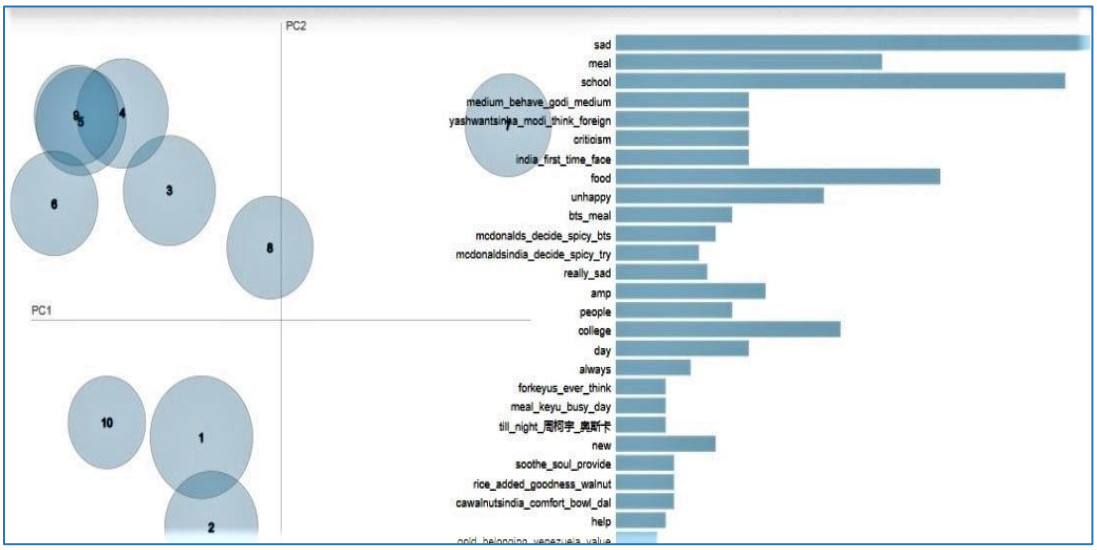

Figure 11 Representation of the top salient terms in the dataset

Each bubble 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 Figure 11, that the words sad, school, and food are the most recurring words in the dataset. Topic models for topics 1, 3, and 5 can be seen in Figures 12, 13, and 14.

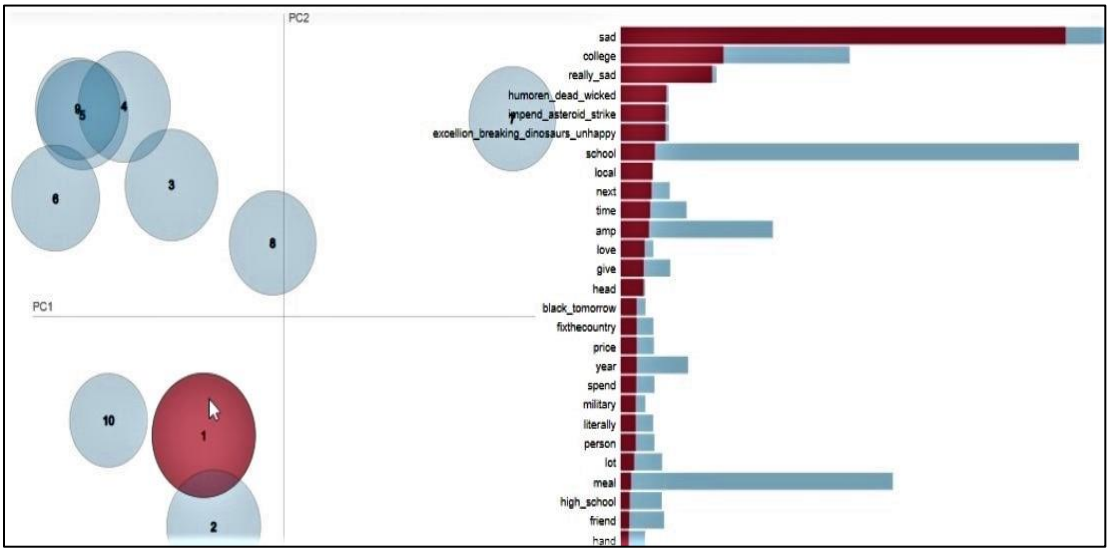

Figure 12 Topic 1's most frequent words (marked in red) the bubble is big hence it means topic 1 is more widespread in the dataset

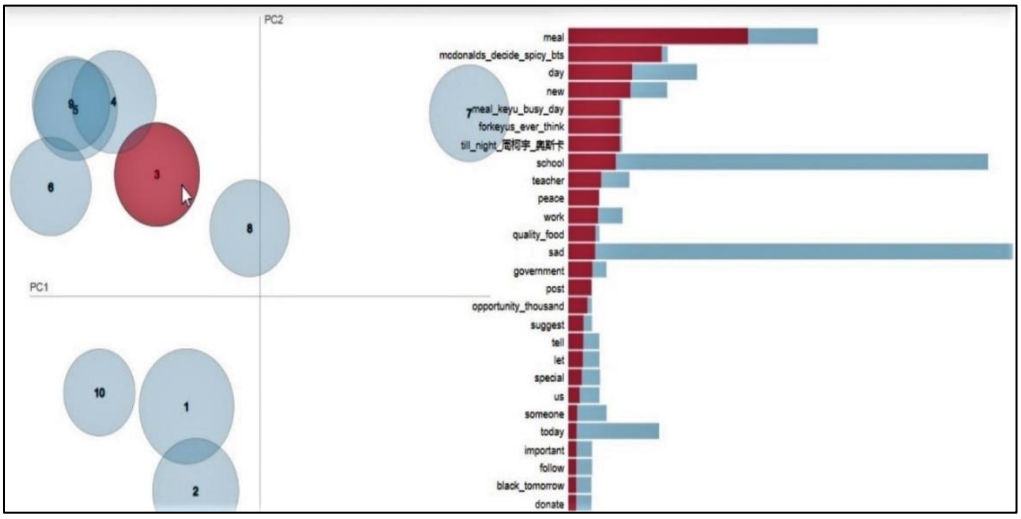

Figure 13 Topic 3's most frequent words (marked in red)

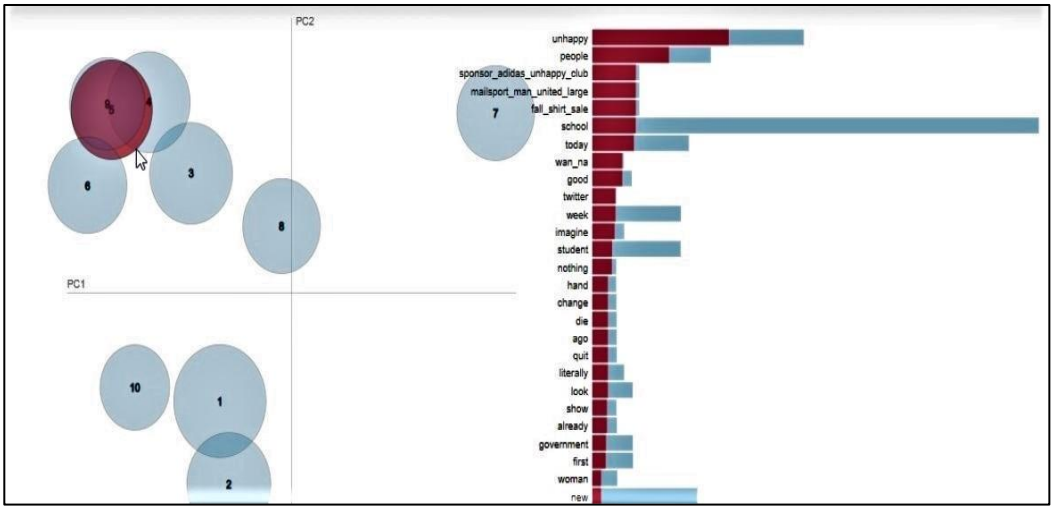

Figure 14 Topic 5's most relevant words

Binary Sentiment Analysis

After entering that we have to enter the path location of the .csv file [31], it will produce an output excel file as shown in Figure 15. The tweets are labeled '0' or '1' where 0 stands for a negative sentiment and 1 stands for a positive sentiment. Figure 16 depicts the evaluation metrics for trained models for sentiment classifiers built using logistic regression, decision tree, and random forest. It can be observed that random forest has the best accuracy. Hence, it can be used in the final framework. The model is also tested for random tweets as shown in Figure 17, the tweet is classified as positive.

	id	tweet	emotion 0
0	31963	#studiolife #aislife #requires #passion #dedication #willpower to find #newmaterialsÃ¢€Â¦	0
1	31964	@user #white #supremacists want everyone to see the new Ã¢€Ã˜ #birdsÃ¢€Â™ #movie Ã¢€Â" and hereÃ¢€Â™s why	0
2	31965	safe ways to heal your #acne!! #altwaystoheal #healthy #healing!!	0
3	31966	is the hp and the cursed child book up for reservations already? if yes, where? if no, when? Ã°ÂŸÂ˜ÂÃ°ÂŸÂ˜ÂÃ°ÂŸÂ˜Â #harrypotter	1
4	31967	3rd #bihday to my amazing, hilarious #nephew eli ahmir! uncle dave loves you and missesÃ¢€Â¦	0
5	31968	choose to be :) #momtips	1
6	31969	something inside me dies Ã°ÂŸÂ¦Ã°ÂŸÂ¿Ã¢Â¨Â¨ eyes ness #smokeyeyes #tired #lonely #sof #grungeÃ¢€Â¦	0
7	31970	#finished#tattoo#inked#ink#loveitÃ¢Â¤Ã¯Â¸Â #Ã¢Â¤Ã¯Â¸ÂÃ¢Â¤Ã¯Â¸Ã¢Â¤Ã¯Â¸Â #thanks#aleeee !!!	1
8	31971	@user @user @user i will never understand why my dad left me when i was so young.... :/ #deep #inthefeels	1
9	31972	#delicious #food #lovelife #capetown mannaepicure #resturantÃ¢€Â¦	0
10	31973	1000dayswasted - narcosis infinite ep.. make me aware.. grinding neuro bass #lifestyle	1
11	31974	one of the world's greatest spoing events #lemans24 #teamaudi	1
12	31975	half way through the website now and #allgoingwell very	0
13	31976	good food, good life , #enjoy and Ã°ÂŸÂ™ÂŒÃ°ÂŸÂÂ•Ã°ÂŸÂ‘Â€Ã°ÂŸÂ‘ÂÃ°ÂŸÂ‘ÂŒÃ°ÂŸÂ‘Â„ this is called ~garlic bread~ ... #iloveitÃ¢	0
14	31977	i'll stand behind this #guncontrolplease #senselessshootings #taketheguns #comicrelief #stillsad	1
15	31978	i ate,i ate and i ate...Ã°ÂŸÂ˜Â€Ã°ÂŸÂ˜Â #jamaisasthi #fish #curry #prawn #hilsa #foodfestival #foodies	0
16	31979	@user got my @user limited edition rain or shine set today!! ! @user @user @user @user	1
17	31980	&#love &#hugs &#kisses too! how to keep your #baby #parenting #healthcare	1
18	31981	Ã°ÂŸÂ‘Â-Ã°ÂŸÂŒÂžÃ°ÂŸÂ‘Â—#girls #sun #fave @ london, united kingdom	0
19	31982	thought factory: bbc neutrality on right wing fascism #politics #media #blm #brexit #trump #leadership >3	1
20	31983	hey guys tommorow is the last day of my exams i'm so happy yay	1
21	31984	@user @user @user #levyrroni #recuerdos memoriesÃ°ÂŸÂ‘Â-Ã¢Â¤Ã°ÂŸÂ”Â—Ã°ÂŸÂ‘Â«Ã°ÂŸÂ‘Â #recuerdos #friends #life #triur	1

Figure 15 Excel file after Sentiment Analysis

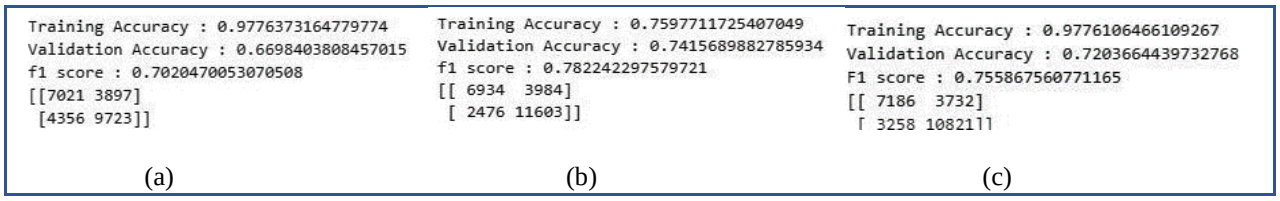

Figure 16 F1 score, Confusion Matrix, Training and Validation Accuracy of **(a)** Logistic Regression **(b)** Decision Tree Classifier (c)Random Forest Classifier

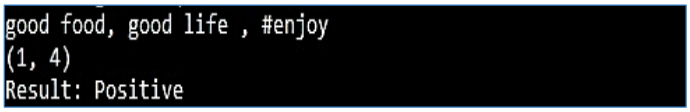

Figure 17 Sentiment Analysis of a tweet

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 [32]. The training of the model and accuracies obtained at different epochs are shown in Figure 18. The emotion classification model built is used in the web portal designed. It can be observed that a training accuracy of 94% is achieved. The evaluation of the model on the test set of 6000 tweets is shown in Figure 19. The figure depicts micro evaluation metrics: accuracy, F1 score, 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 20-25 indicate the
classification of the emotion for the post given as input.

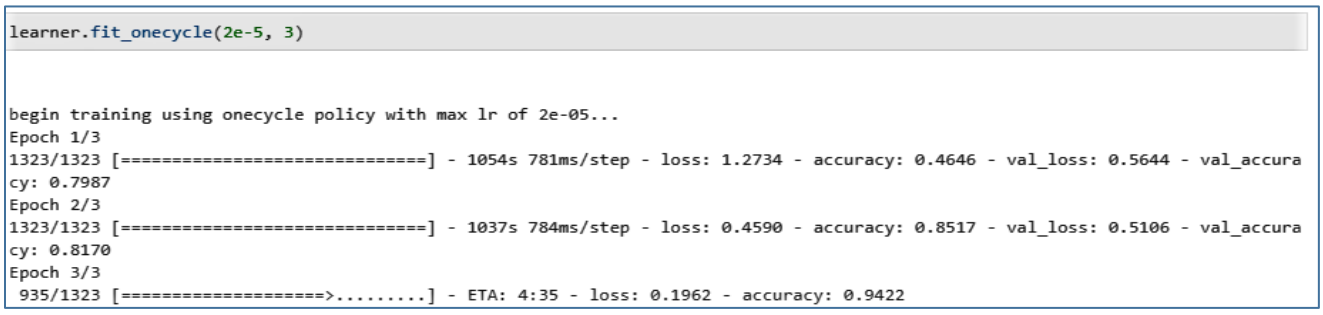

Figure 18 Bert model created with an accuracy of 94 percentage

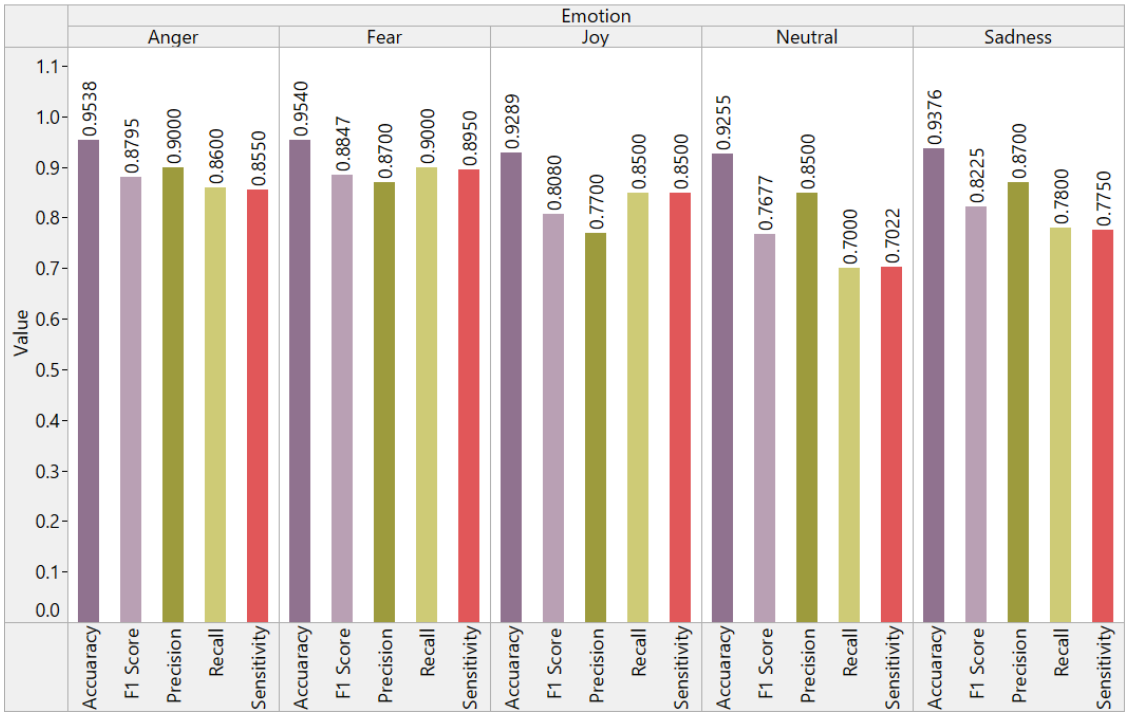

Figure 19 Performance evaluation of test set using BERT model

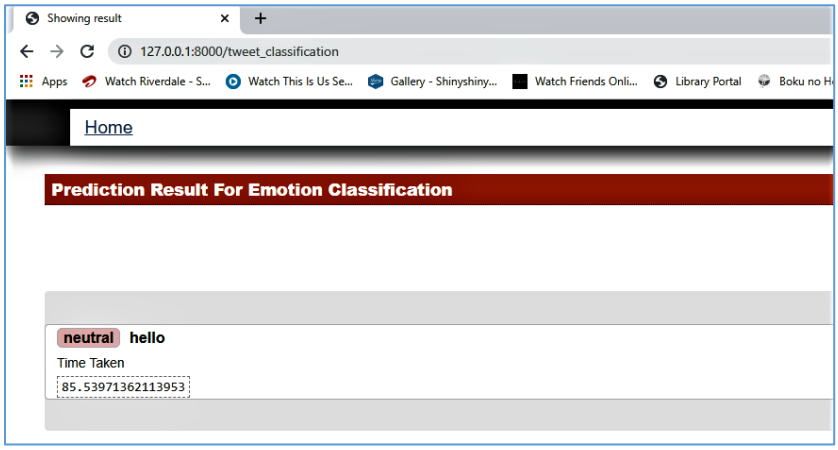

Figure 20 Neutral Tweet with the time taken 85 seconds

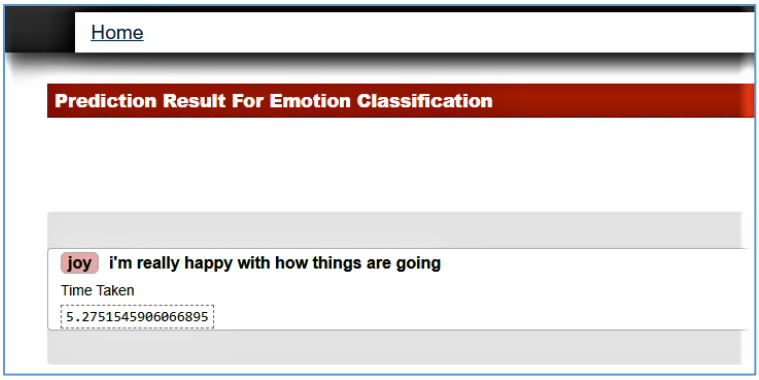

Figure 21 Joy Tweet with the time taken 5 seconds

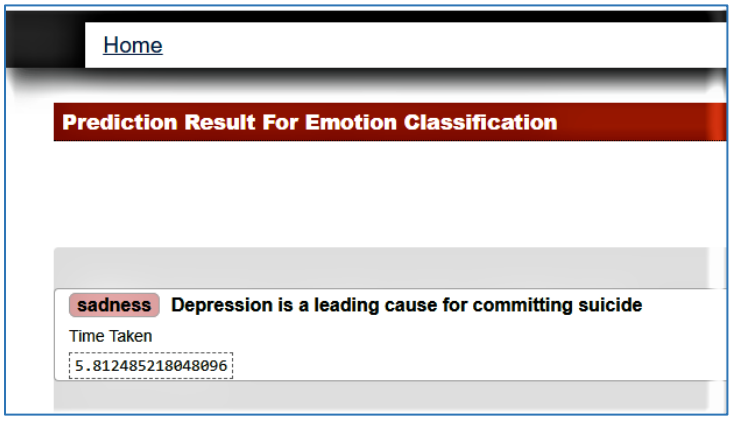

Figure 22 Sadness Tweet with the time taken 5 seconds

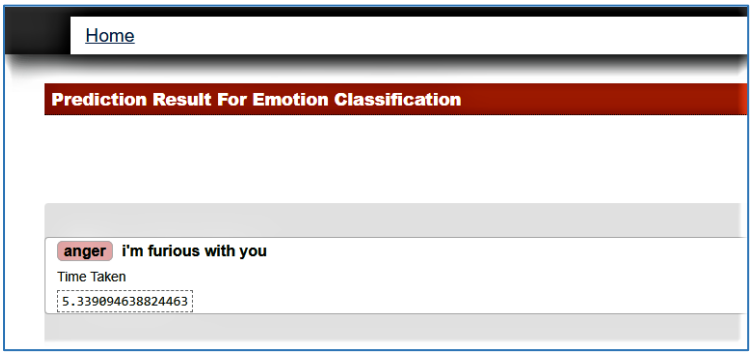

Figure 23 Anger Tweet with the time taken 5 second

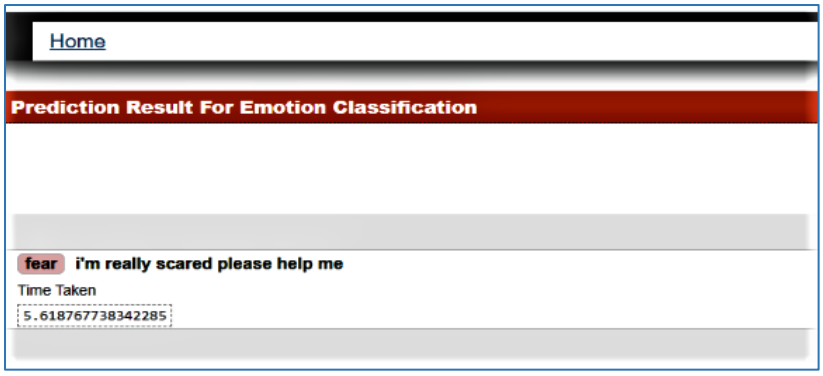

Figure 24 Fear Tweet with the time taken 5 seconds

CONCLUSION

Sentiment analysis is an area of learning for examining opinions expressed in the text on numerous social media websites. Our projected model used plentiful algorithms to enhance the precision of categorizing tweets as a positive or negative sentiment. In our paper, we offered an outline for detecting user's mental stress from twitter's social media facts and figures. Using real-life social media data as the foundation, I calculated diverse relations between user's psychological stress states and their social interaction manners. To completely leverage twitter's material on users' tweets, I projected a model which uses Random Forest Algorithm to make predictions. 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. Random Forest Classifier was a better algorithm due to a high accuracy which was 97.78%. In this paper, we also emphasized the importance of data visualizations as it helps us in getting an apt understanding of our data. In this paper, we used the helpful pre-trained BERT model and fine-tuned it for the sentiment classification task of 5 different classes – Joy, Sadness, Neutral, Anger, Fear. Even without a complicated architecture, our model successfully outdid intricate architectures like recursive, recurrent, and convolutional neural networks. Hence, we have verified the classification competence in NLP supported by deep contextual language models like BERT. We also executed Topic Modelling which has witnessed a lot of acclaim in past years and is used in a variety of uses, including the examination of a plentiful of news articles, tag assignment for any document, and search interfaces that are centered on topics. In this suggested work, the implementation of Latent Dirichlet Allocation is used to analyze the data. Extracts are analyzed by using LDA to settle on the number of topics and outline the percentage of a word in a specific topic. The outcomes presented that the extracted topics display a significant structure in the data. The suggested method gives accurate results to any dataset required.

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 also 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 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 of these features combined would help develop an astounding analyzing tool.

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

DECLARATION

Ethics approval and consent to participate

NOT APPLICABLE

Consent for publication

NOT APPLICABLE

Availability of data and materials

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

Funding

NOT APPLICABLE

Competing interests

NOT APPLICABLE

Authors' contributions

First author: Author has designed and implemented all the phases of the project

Second author/Corresponding author: Author has guided throughout the process of project and has performed result analysis.

Third author: Author has finalized the structure and content of the manuscript and has done the proof reading.

Acknowledgements: NIL

Authors' information:

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 PhD 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 Assistant Professor in the Department of Electronics and Communication Engineering. He is a senior member of IEEE and his areas of interests include Signal processing, Pattern classification, and Machine learning.

References

1. B. Wang, Y. Liu, Z. Liu, M. Li, and M. Qi, "Topic selection in latent Dirichlet allocation," 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014, pp. 756-760, doi: 10.1109/FSKD.2014.6980931.
2. B. Liu and L. Zhang, A Survey of Opinion Mining and Sentiment Analysis. Boston, MA: Springer US, 2012, pp. 415–463.
3. M. Munikar, S. Shakya and A. Shrestha, "Fine-grained Sentiment Classification using BERT," 2019 Artificial Intelligence for Transforming Business and Society (AITB), 2019, pp. 1-5, doi: 10.1109/AITB48515.2019.8947435.
4. Pak, Alexander & Paroubek, Patrick. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of LREC. 10.
5. Agarwal, Apoorv & Xie, Boyi & Vovsha, Ilia & Rambow, Owen & Passonneau, Rebecca. (2011). Sentiment Analysis of Twitter Data. Proceedings of the Workshop on Languages in Social Media.
6. M. K. Peddinti and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter," in Analyzing Microtext Workshop, AAAI, 2011.
7. Davidov, Dmitry & Tsur, Oren & Rappoport, Ari. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2. 241-249.
8. P. Anupriya and S. Karpagavalli, "LDA based topic modeling of journal abstracts," 2015 International Conference on Advanced Computing and Communication Systems, 2015, pp. 1-5, doi: 10.1109/ICACCS.2015.7324058.
9. Si, Xiance & Sun, Maosong. (2008). Tag-LDA for Scalable Real-time Tag Recommendation. Journal of Computational Information Systems. 6.
10. Krestel, Ralf & Fankhauser, Peter. (2012). Personalized topic-based tag recommendation. Neurocomputing. 76. 61-70. 10.1016/j.neucom.2011.04.034.
11. M. e. a. Peters M, Neumann M, "Deep contextualized word representations," 2018.
12. S. T. S. I. Radford A, Narasimhan K, Improving language understanding by generative pre-training, 2018.
13. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423
14. Z. Jin, X. Lai and J. Cao, "Multi-label Sentiment Analysis Base on BERT with modified TF-IDF," 2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN), 2020, pp. 1-6, doi: 10.1109/ISPCE-CN51288.2020.9321861.
15. M. Zubair, K. Aurangzeb, A. Shakeel, Q. Maria, K. I. Ali, and Z. Quan, "Lexicon-enhanced sentiment analysis framework using rulebased classification scheme," Plos One, vol. 12, no. 2, p. e0171649, 2017.
16. D. Zeng, Y. Dai, F. Li, J. Wang, and A. K. Sangaiah, "Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism," J. Intell. Fuzzy Syst., vol. 36, no. 5, pp. 3971–3980, 2019. [Online]. Available: https://doi.org/10.3233/JIFS-169958
17. R. Socher, A. Perelygin, and J. Wu, "Recursive deep models for semantic compositionality over a sentiment treebank," Proc. Conf. Empir. methods Nat. Lang. Process., vol. 1631, p. 1642, 2013.
18. Z. Wang, R. S. M. Goh, and Y. Yang, "A method and system for sentiment classification and emotion classification," Patent Cooperation Treaty (PCT) Application, PCT/SG2015/050469, 2014.
19.] Z. Wang, R. S. M. Goh, and Y. Yang, "SentiMo-A Method and system for fine-grained classification of social media sentiment and emotion patterns," Singapore Patent Application10201407766R, 2014.
20. Z. Wang and J. C. Tong, "ChiEFS-A method and system for Chinese hybrid multilingual emotion fine-grained sensing of text data," Singapore Patent Application No. 10201601413Q, 2015.
21. Go, Alec & Bhayani, Richa & Huang, Lei. (2009). Twitter sentiment classification using distant supervision. Processing. 150. (n)
22. Balahur, Alexandra & Steinberger, Ralf & Kabadjov, Mijail & Zavarella, Vanni & van der Goot, Erik & Halkia, Matina & Pouliquen, Bruno & Belyaeva, Jenya. (2013). Sentiment Analysis in the News. Proceedings of LREC. (n-1)

23. Jonathon Read. 2005. 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]
24. Boiy, E., Moens, MF. A machine learning approach to sentiment analysis in multilingual Web texts. *Inf Retrieval* **12,** 526–558 (2009). https://doi.org/10.1007/s10791-008-9070-z [N+1]
25. Fangtao Li, Minlie Huang, and Xiaoyan Zhu. 2010. Sentiment analysis with global topics and local dependency. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI Press, 1371–1376. (N+2)
26. Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in IEEE Access, vol. 5, pp. 2870-2879, 2017, doi: 10.1109/ACCESS.2017.2672677.
27. S. Pradha, M. N. Halgamuge and N. Tran Quoc Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-8, doi: 10.1109/KSE.2019.8919368.
28. D. Deepa, Raaji and A. Tamilarasi, "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, pp. 786-790, doi: 10.1109/I-SMAC47947.2019.9032456.
29. S. Chaturvedi, V. Mishra and N. Mishra, "Sentiment analysis using machine learning for business intelligence," 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017, pp. 2162-2166, doi: 10.1109/ICPCSI.2017.8392100.
30. J. Ho, D. Ondusko, B. Roy and D. F. Hsu, "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, pp. 1066-1071, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00191.
31. Keerthi Vardhanapu, May 14, 2020, "Sentiment analysis", IEEE Dataport, doi: https://dx.doi.org/10.21227/e2aq-xv12.
32. Damian, March 11, 2021, “Detecting Emotions in Text”, https://data.world/damof/detecting-emotions-in-text