Sentiment Analysis on National Education Policy Change 2020

Ritika Biswas[1], Naman Vyas[2], Dr. M. Baskar[3]

[1]Department of computer science and engineering, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603203
[2]Department of computer science and engineering, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603203
[3]Department of computer science and engineering, SRM Institute of Science and Technology, Kattankulathur, Chengalpattu, Chennai, Tamilnadu, India-603203
[1]1999.ritikabiswas@gmail.com,[2] namanvyas786@gmail.com, [3]baashkarcse@gmail.com*

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Examining sentiment is a cycle to recognize the assessment of a text. Individual's write reviews in web-based media referencing their understanding related to an occasion and are likewise intrigued to know other's insight on a similar occasion. This categorization can be accomplished utilizing SA. SA extracts structure less text reviews related to an item surveys, an occasion, and so on, from all reviews written by various clients and groups the reviews into various classifications as one or the other positive or negative or impartial assessment. This is otherwise called polarity classification. Individuals these days use emojis in their content progressively to communicate their emotions or reiterate their words. Prior AI methods just include the order of text, emojis or pictures exclusively where emojis with text have consistently been dismissed, accordingly overlooked heaps of feelings. This exploration proposed a calculation and strategy for estimation examination utilizing both text and emojis. In this work, information was investigated utilizing deep learning calculation to find sentiments from tweets utilizing a few highlights like Term frequency inverse document frequency, emoji vocabularies, N-gram and Bag of words.

Keywords—Sentiment Analysis, LSTM, Deep Learning, Emoticon Classification, Twitter data, NEP-National Education Policy

1. Introduction

Categorizing sentiment is a cycle to discover the assessment of a text. Breaking down sentiment is an interaction of knowing client feelings for a specific thing which might be an occasion or subject or individual of late patterns. Twitter stage utilizes a tweet which is in sentence structure to indicate conclusions. The objective is to figure the sentiment exactness of sentences extricated from the text obtained from twitter comments.SA should be possible for twitter information which will arrange the tweet as one or the other positive or negative. This classification assists required associations with discovering assessments of individuals about their item, occasions, and much more from the tweets. The most difficult part is evaluating a word in SA. An assessment word might be negative/positive contingent upon different circumstances.

The time of getting significant experiences from web-based media information has now shown up with the development in innovation. Customarily, investigation of sentiment has been done on text; be that as it may, a lot of information is presently being transferred as audits, pictures, emojis, and recordings. By assessing this information, the estimation of general society toward a particular matter can be broke down, analyzed and found. Throughout the long term, individuals thought about emojis as a mode of correspondence that is utilized in writings or exclusively to devote one's assumption in a productive way.

Emojis are emblematic articulations comprising this sort of tokens, for example, \:", \=", \-", \)", or \(" also, usually address outward appearances. We can read emojis horizontally also, as \:- ("(miserable expression), or \(ˆˆ)" (glad expression) and numerous different ways to express an emotion. Checking these images of emojis alongside the text is colossally important to get the genuine suppositions like, satisfaction, disappointment, outrage, misery, and so on which are at that point arranged in certain, negative and nonpartisan. The overall research of sentiment analysis (SA) manages either emojis or text.

Greatest explores in sentiment analysis have been performed only on text collected through social websites utilizing machine learning (ML) algorithms. Be that as it may, SA on both content and emoji has been generally overlooked because of the absence of assets and intricacy of emojis. Text classification is one of the best explorable zones as it removes sentiment utilizing diverse machine learning what's more, deep learning methods with the assistance of later innovations. Yet, research shows deep learning was seldom executed in sentiment analysis on both emoji and text information mix. Accordingly, this examination has dissected the content and emojis in blend to discover the notions. Likewise, the exploration built up an emoji vocabulary, examined the conclusions by applying emoji vocabularies alongside some content highlights such as Term frequency inverse document frequency, N-gram, and bag of words using deep learning algorithm.

2. Related Works

We have gone through some of past works in the related fields for knowledge gaining and inspiration. Nowadays, sentiment analysis can be performed on different domains and languages, as in[1].

In [2], they used text sentiment analysis, visual sentiment analysis. The geo-sentiment error among information objects in a nearby geological zone, and noticing assorted sentiments from sight and sound information objects was addressed. The extricated sentiments were accumulated geologically for the motivation behind extricating more exact neighbourhood local experiences.

In [3], they tried to predict the level of teaching performance automatically by collecting data through students' text feedback and using lexicon-based approach. In [4], they utilized LDA that extricates subject of archives where the subject is addressed as the presence of the words with various point probability. They tracked down that the best blend and also the values of alpha, beta, required points, edge and perplexity.

In [5], Word2Vec and ISODATA clustering algorithm was proposed. In [6], they enhanced the exactness of the SA results by combining functions from two sentiment lexica. Detection of sentiment-bearing terms and negative sentiments was performed well by using a principal lexicon, so they used that respective lexicon. Then to order the remainder of the information, they utilized a second lexicon. Experimental outcomes show that this technique enhanced the exactness of the outcomes while being compared to using only one lexicon.

In [7][21], they separated the emoticons out of the content, regardless of whether it is a negative, neutral or positive emotion. In [8], they have used opinion mining, ML, NLP, dataset labelling and sentiment polarity.

The sentiment analysis in Bengali language is restricted to only micro-blogging and Bengali corpus. So, in [9] they have focused on an exceptional area that is Bangladesh Cricket where people's point of view in Bengali languages on social website is expressed every moment.

In [10], they proposed the strategy to naturally examine Thai sentiment of purchaser's survey in the item, cost, and delivery dimensions by utilizing a multi-dimensional dictionary and sentiment remuneration method. A customer's survey in the Thai language is tokenized utilizing the longest coordinating with the algorithm. At that point, it is broke down to discover its sentiment. Sentiment remuneration method is utilized to consequently repay the sentiment to a dimension where the shopper's audit specifies the sentiment without a dimension.

In [11], the SentiStrength algorithm targets identifying sentiment polarity and strength of a book. Accordingly, it assists with distinguishing improper full of feeling expressions, related to segregation and animosity. The discoveries uncover a winning solid negative sentiment of the investigated tweets.

In [12], they used Naïve Bayes Classifier for classification and the experiment was divided into two parts, using stemming and without stemming. In [13], for recognizing the sentiments, they have utilized a backtracking algorithm, where the core of this methodology is a sentiment dictionary. Also, the examination showed the backtracking algorithm performed more than 70% accuracy to identify genuine public sentiment.

In [14][22], an endeavour was made to propose an examination strategy for the SA of the Twitter dataset. In this process the polarity of each and every tweet was calculated to recognize whether the tweet is positive or negative. A sentiment polarity is the feelings of the client like furious, pitiful, cheerful, and satisfied. The proposed component has been carried out in Python.

In [15], they suggested a creative sentiment investigation strategy in light of the presence of mind information. They made their unique Oman travel industry philosophy in light of ConceptNet. Substances are recognized from the tweets utilizing Part-of-Speech tagging and elements are contrasted and ideas in the sector explicit cosmology. After this, the sentiment of the extricated substances is dictated by the consolidated sentiment vocabulary approach. At last semantic directions of domain explicit highlights are joined.

In [16], the paper investigates the field of Natural Language Processing (NLP) for the Sentiment Analysis of Chinese textual data. They proposed a DL strategy where they used sentence-based approach in the SA of online audits to gain greater granularity and escalate classification precision.

In [17], a comparison was completed using the K-Nearest Neighbour (K-NN),SVM and naïve bayes strategies along with RapidMiner which resulted in a NB exactness estimation of 75.58%, Support Vector Machine exactness estimation of 63.99% and K-Nearest Neighbors precision of 73.34%.

In [18], they proposed to utilize Partial Textual Entailment for estimating semantic likeness between the tweets to bunch comparable tweets together. The technique is expected to lessen the weight of the sentiment analyser and make the processing quicker. Besides, we additionally propose an adjustment in a current strategy

for Partial Textual Entailment which can be additionally received for some Natural Language Processing applications.

In [19], they developed a danmaku sentiment word reference and present another technique utilizing sentiment word reference and Naïve Bayes for the sentiment investigation of danmaku audits. The technique is extraordinarily useful in managing the general enthusiastic direction of a danmaku video, what's more, anticipating its prevalence. Through the cycles of extricating enthusiastic data from a danmaku video, characterizing sentiment, and envisioning information, the time circulation of the seven sentiment dimensions can be acquired. What's more, a weighted count can be led for arranging the sentiment polarity of danmaku surveys. Exploratory outcomes show that the proposed technique significantly affects sentiment score and polarity discovery.

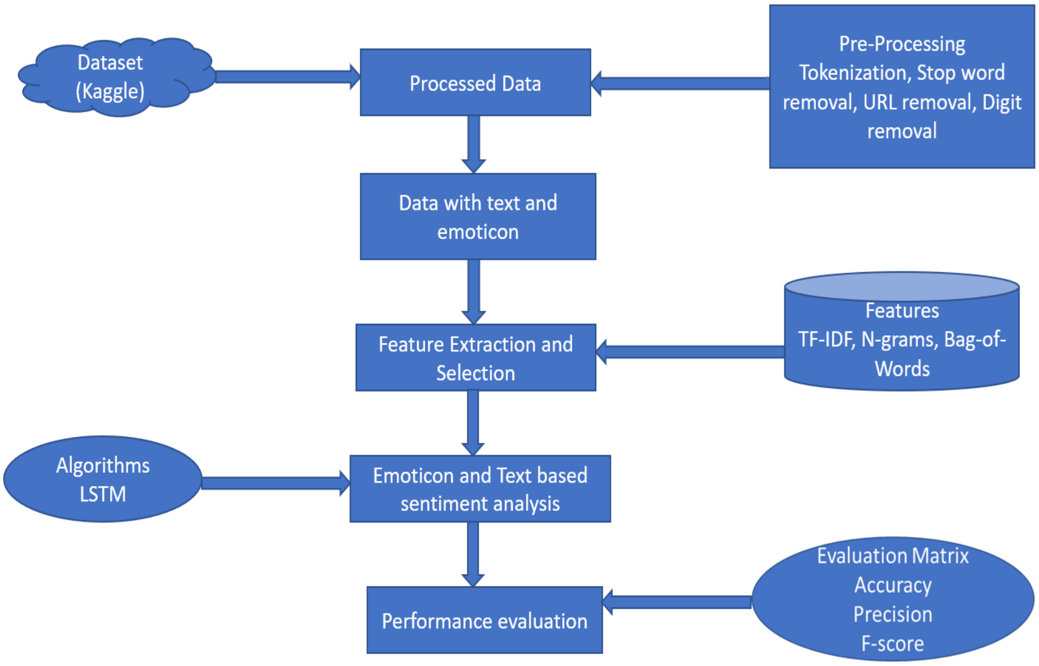

Dataset
(Kaggle)
Processed Data
Pre-Processing
Tokenization, Stop word removal, URL removal, Digit removal
Data with text and emoticon
Features
TF-IDF, N-grams, Bag-of-Words
Feature Extraction and Selection
Algorithms
LSTM
Emoticon and Text based sentiment analysis
Evaluation Matrix
Accuracy
Precision
F-score
Performance evaluation

Figure 1: Architecture Diagram

In [20], the work targets contrasting the presentation of various AI techniques in performing sentiment examination of information collected through twitter. The proposed technique utilizes term recurrence to discover the orientation of the expressed sentiment. The presentation of Logistic regression, support vector machine, and Multinomial Naive Bayes, algorithms in sentence order was looked at.

3. Proposed Work

Sentiment analysis is consistently developing subset of NLP. There are many researchers who have solved the issues of sentiment analysis from text, emoji, pictures, what's more, sound or recordings independently. Not many investigates have been done on emoji for discovering sentiments. Additionally, related works segment conveys there is a degree for additional expansion in the area of sentiment analysis with both emoji and text. Accordingly, the targets of this research are:

- SA utilizing both (emojis and text) via online media information.

- Creating emoticon lexicon.
- To upgrade the arrangement precision of SA by utilizing a DL calculation.
- The principal intention is to comprehend public sentiments with respect to the new education policy.
- To comprehend whether students are agreeable to the adaptability and selection of subjects, multilingualism-based move, tech-based training.
- To understand whether the students are pleased to the new methodology, do they believe that it's helpful for themselves or not.
- Since going through all studies is drawn-out, consequently the prerequisite for robotization to gauge, constantly separate and orchestrate the sentiments.

IMPLEMENTATION

4.1 Dataset:

The information for this exploration was gathered from kaggle. This dataset contains tweets identified with the new education strategy 2020. Every one of the tweets is in the English language. The tweets are from 31 July 2020 to 12 August 2020. The dataset contains 18200 tweets.

4.2 Preprocessing:

In this step, the information gathered was tokenized, and the stop words, URL's, and digits were eliminated, however punctuation and emoticons were not eliminated. This pre-prepared information was the used to find sentiments with emoticons.

- Tokenization- Tokenization is the process of converting text into tokens before transforming it into vectors. Tokenization is a typical assignment in Natural Language Processing (NLP). It's an essential advance in both conventional NLP strategies like Count Vectorizer and Advanced Deep Learning-based models like Transformers. This is a process of separating a piece of text into more modest units called tokens. Here, tokens can be words, characters, or subwords. Moreover, tokenization can be classified into three kinds –subword (n-gram characters), character and word tokenization.

- Removing stop words - A stop word is an ordinarily used word, (for instance, "a", "the", "in", "an") that a search engine has been modified to ignore, both when ordering passages for looking and while recovering them as the consequence of a search inquiry. These words are not needed to occupy space in our information base, or occupying important processing time. To implement this, we can eliminate them effectively, by eliminating a rundown of words that are considered stop words. Natural Language Toolkit (NLTK) in python has a rundown of stopwords listed in 16 unique languages.

- URLs – Dispose of these URLs by means of regular expression matching.

- @username – Eliminate "@username" through regex coordinating or supplant it with nonexclusive word AT_USER.

- #hashtag – Supplant hashtags with precisely the same word without the hash.

- Punctuations and additional white spaces – Eliminate punctuation toward the beginning and finishing of the tweets. Additionally, supplant numerous whitespaces with a solitary whitespace.

4.3 Feature Extraction and Selection:

One significant advance in building a classifier is choosing what features of the information are important, and how to encode those highlights. Likewise, we can utilize the presence or nonappearance of words that show up in tweet as highlights. In the preparation information, we can part each tweet into words and add each word to the element vector. A portion of the words probably won't show the sentiment of a tweet and we can sift them through. At that point consolidate singular component vector into an enormous rundown that contains every one of the highlights and eliminate copies in this rundown.

In this step, Term frequency inverse document frequency, N-grams, Bag-of-Words and emoticon lexicons were chosen.

- N-GRAMS - A N-gram model is worked by counting how often word sequences happen in content and then estimating the probabilities.

- TF-IDF - Short for term frequency-inverse document frequency is a numerical measurement that is designed to show how essential a word is to a document in a collection.

- Bag-of-Words –This technique simply makes a bunch of vectors which contains the count of word occurrences in a document (surveys).

4.4 Algorithm:

The main algorithm used for classifying both text and emoticon is the deep learning algorithm – Long Short Term Memory (LSTM). In this process, we have utilized separated 10 -fold cross validation to enhance the precision. The dataset was divided into 10 sections, among them, 9 sections were used for training the data and 1 was used for testing the data. This method was continued for all the training-testing parts combination.

While executing LSTM, firstly we have to decide what information is to be removed from the cells, which is decided by the ‘forget gate layer’. Secondly, we have to store the required information in the cells by updating the values and creating a vector of those new values.

After this, old cells are replaced by the new cells. Finally, the output has to be decided which will be a filtered version of the LSTM cells, obtained using sigmoid layer.

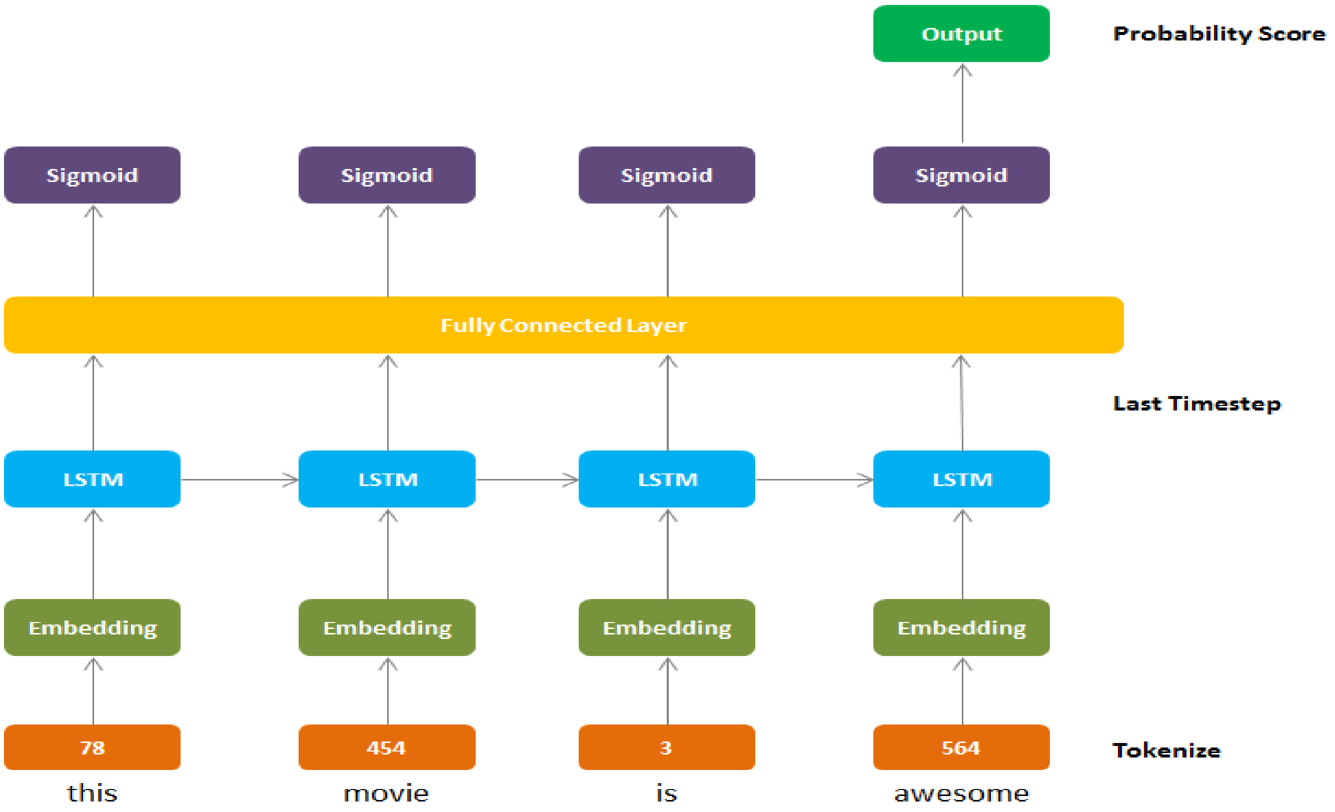


FIGURE 2: UML DIAGRAM

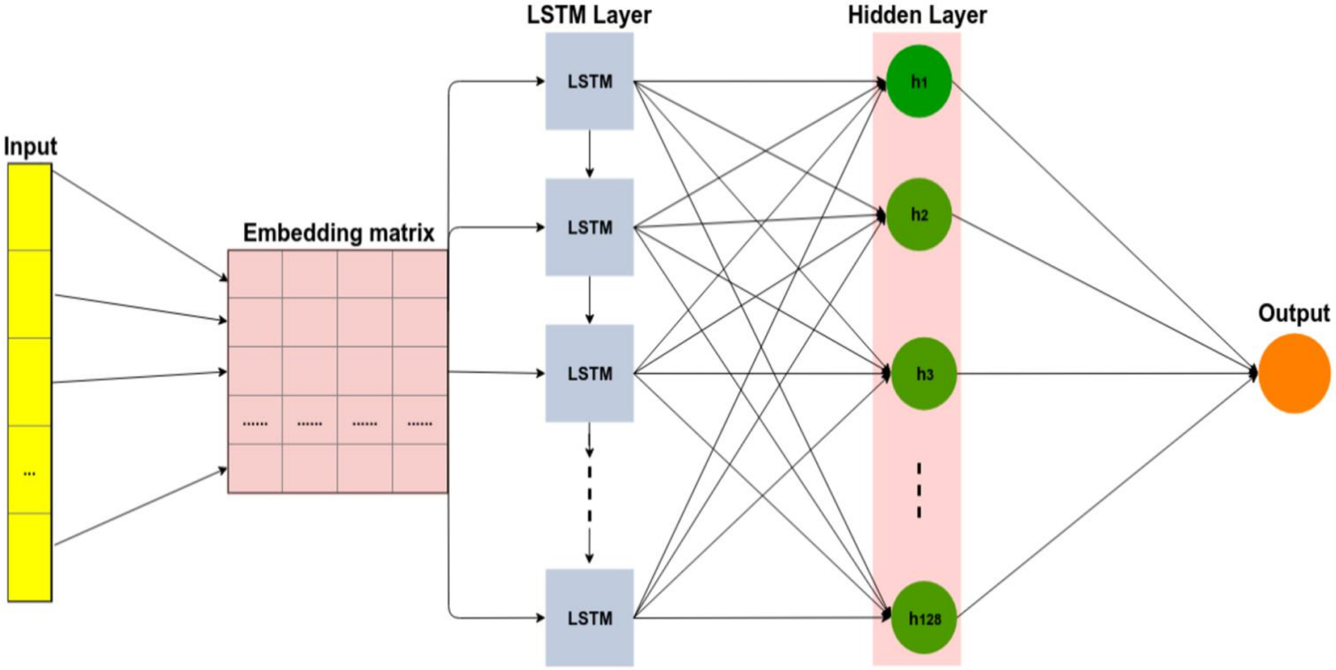


Figure 3: Graph Representation

Algorithm:

Input: education_policy_dataset

Output: Sentiment (Positive/Negative)

Notation: Feature-F, Deep Learning-DL, TF-IDF-TI, Bag of words-BOW, n-gram-ng, emoticons-E

```
Begin
        file = url, tokenize, digit removing ,stop words
        F1- extract the features, TI, BOW, ng and E from file
        Find the score by DL algorithms
                for each F1 in file do
                    while i=1…n do
                    while j=1…n do
        DL_score[i][j] = score applying DL;
                    end while
                     end while
               end for
        end
```

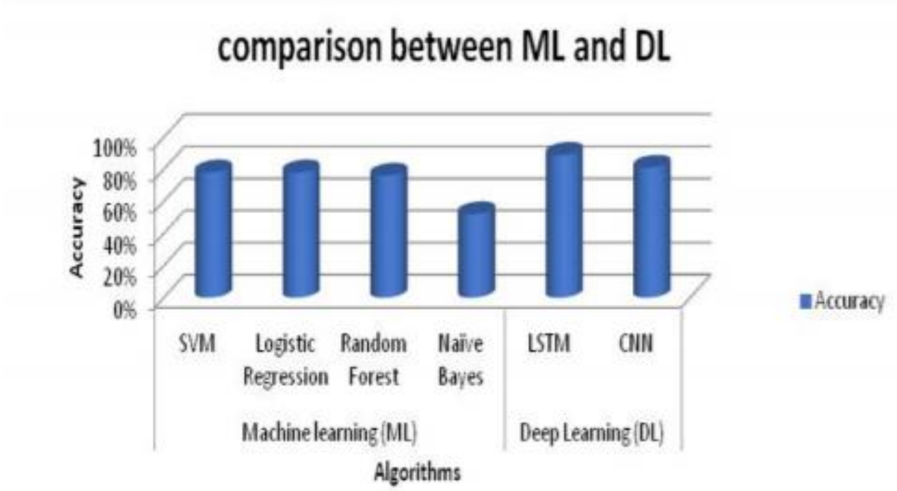
4. Result and Discussion:

In this research, we have used the programming language – 'python' for the analysis. From the results we have obtained it is apparent that, analyzing text and emoticon with the right feature selection and updated emoticon lexicons as stated before is better than analyzing only text. However, both of these methods vary slightly around 1–3% based on their performance. The categorization exactness of both Support Vector Machine and logistic regression wins to be superior when compared to machine learning classifiers, for example, Naïve Bayes and Random Forest. Generally, the performance of Naïve Bayes was foundto be exceptionally unfit for this experiment. While, for the situation of deep Learning, Long Short Term Memory shows the preferable classification exactness over CNN. In general, DL algorithms are outperforming ML algorithms. DL accomplished elevated categorization exactness because it tackled the issue point to point and naturally extricates highlights.

The performance of DL algorithms, for example, Long Short Term Memory and Convolutional Neural Network was better when compared to all the Machine Learning algorithms with 89% and 81% precision respectively for combined text and emoticon. Both of these algorithms also worked better for categorizing only textual data. There was a difference of 1-9% in precision. Although, this model had an issue of 'overfitting' which was settled by acquiring different tuning techniques of hyper-boundary, for example, diminish the network's ability (RNC), regularization (R), and dropout layers (DO). Nonetheless, extremely negligible refinement is accomplished. Overfitting seems to be an issue in the R model as it begins along with the baseline model.

	Algorithms	Accuracy
Machine Learning (ML)	Naïve Bayes	52%
	Logistic Regression	76%
	Random Forest	78%
	SVM	78%
Deep Learning (DL)	Long Short Term Memory	89%
	Convolutional Neural Network	81%

Table 1: Comparison between machine and deep learning on accuracy



Graph 1:Graph representation of table 1 comparison

5. Conclusion

This paper concluded a technique for classifying sentiments using both text and emoticon from social media websites. The respective social media website used here was twitter and data were gathered related to the national education policy change 2020.This research also showed the improved accuracy because of considering emoticons while analyzing sentiments. The main algorithm used here was a deep learning algorithm – Long short term memory (LSTM).The general outcome indicated that considering emojis along with text clearly affected the SA. Additionally, it is established that Deep Learning algorithms performs better compared to Machine Learning algorithms. Finally, existing methods were outperformed. For future improvement, this exploration can be broadened to the area where more than one language is used in the dataset.

REFERENCES

a. Lama Hamandi, Sabra, Mohamad A. El Abed, Rached N. Zantout, (2017), Sentiment Analysis: Arabic Sentiment Lexicons, IEEE, DOI: 10.1109/SENSET.2017.8125054

b. Cyrus Shahabi, Seon Ho Kim, Abdullah Alfarrarjeh, Sumeet Agrawal, (2017), Geo-spatial Multimedia Sentiment Analysis in Disasters, IEEE, DOI: 10.1109/DSAA.2017.77

c. Nyein NyeinMyo, KhinZezawarAung , (2017), Sentiment Analysis of Students' Comment Using Lexicon Based Approach, IEEE, DOI: 10.1109/ICIS.2017.7959985

d. RetnoKusumaningrum, Mohammad F. A. Bashri, (2017), Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity WordcloudVisualization, IEEE, DOI: 10.1109/ICoICT.2017.8074651

e. Qingsong Yu, Xiaobo Zhang, (2017), Hotel Reviews Sentiment Analysis Based on Word Vector Clustering, IEEE, DOI: 10.1109/CIAPP.2017.8167219

f. Riza Batista-Navarro , Khaleel Malik, , Maria Sharmina , Victoria Ikoro (2018), Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers, IEEE, DOI: 10.1109/SNAMS.2018.8554619

g. Meena Belwal ,Satuluri Vanaja, (2018), Aspect-Level Sentiment Analysis on E-Commerce Data, IEEE, DOI: 10.1109/ICIRCA.2018.8597286

h. Oludayo O Olugbara, Kudakwashe Zvarevashe, (2018), A Framework for Sentiment Analysis with Opinion Mining of Hotel Reviews, IEEE, DOI: 10.1109/ICTAS.2018.8368746

i. MahfuzurRahaman, Nazmul Islam , Shamsul ArafinMahtab , (2018), Sentiment Analysis on Bangladesh Cricket with Support Vector Machine, IEEE, DOI: 10.1109/ICBSLP.2018.8554585

j. ChayapolMoemeng, Paitoon Porntrakoon, (2018), Thai Sentiment Analysis for Consumer's Review in Multiple Dimensions Using Sentiment Compensation Technique (SenseComp), IEEE, DOI: 10.1109/ECTICon.2018.8619892

k. Olena Levchenko, Marianna Dilai, (2018), Discourses Surrounding Feminism in Ukraine: a Sentiment Analysis of Twitter Data, IEEE, Volume: 2, DOI: 10.1109/STC-CSIT.2018.8526694

l. Novanda Alim SetyaNugraha, Muhammad ZidnyNaf'an, Ezar Mega Risondang, AfiatariLarasati, AlhamdaAdisokaBimantara, (2019), Sentiment Analysis of Cyberbullying on Instagram User Comments, Reasearch Gate, DOI: 10.21108/JDSA.2019.2.20

m. TapasyRabeya, Narayan Ranjan Chakraborty, Sanjida Ferdous, Manoranjan Dash, Ahmed Al Marouf, (2019), Sentiment Analysis of Bangla Song Review- A Lexicon Based Backtracking Approach, IEEE, DOI: 10.1109/ICECCT.2019.8869290

n. Priyanka Chauhan, Sanjeev Dhawan, (2019), Sentiment Analysis of Twitter Data in Online Social Network, IEEE, DOI: 10.1109/ISPCC48220.2019.8988450

o. T.Meyyappan, Vallikannu Ramanathan, (2019), Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism, IEEE, DOI: 10.1109/ICBDSC.2019.8645596

p. Yan Pang, Denis Zuba, Jun Sheng, Lee, (2019), Sentiment Analysis of Chinese Product Reviews using Gated Recurrent Unit, IEEE, DOI: 10.1109/BigDataService.2019.00030

q. ApriandyAngdresey, MeylanWongkar, (2019), Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter , IEEE, DOI: 10.1109/ICIC47613.2019.8985884

r. Manpreet Kaur, SachinLakra, Shailja Gupta, (2019), Sentiment Analysis using Partial Textual Entailment, IEEE, DOI: 10.1109/COMITCon.2019.8862241.

s. Guanghao Jin, Zhi Li, Rui Li, (2020), Sentiment Analysis of Danmaku Videos Based on Naïve Bayes and Sentiment Dictionary, IEEE, Volume: 8, DOI: 10.1109/ACCESS.2020.2986582.

t. Poornima. A,K. Sathiya Priya, (2020), A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques, IEEE, DOI: 10.1109/ICACCS48705.2020.9074312.

u. M .Baskar, J. Ramkumar,Ritik Rathore, Raghav Kabra, "A Deep Learning Based Approach for Automatic Detection of Bike Riders with No Helmet and Number Plate Recognition", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1844-1854, ISSN: 2005-4238, April 2020.

v. M .Baskar, J. Ramkumar, V.Venkateswara Reddy, G.Naveen Reddy, “Cricket Match Outcome Prediction using Machine Learning Techniques”, International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp: 1863-1871, ISSN: 2005-4238, April 2020.