International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Socialomics of Malaria with Topic Modelling

Navina Bane
Department Of Computer Science
University of Mumbai
Mumbai, India
navinabane12345@gmail.com

Renuka Patle
Department Of Computer Science
University of Mumbai
Mumbai, India
renukapatle2@gmail.com

Elizabeth Leah George
Asst.Professor, Dept Of CIS
Nagindas Khandwala College
Mumbai, India
elizabeth@nkc.ac.in

Dr. Jyotshna Dongardive
Head, Dept Of Computer Science
University of Mumbai
Mumbai, India
jyotss.d@gmail.com

Abstract

New avenues for social science study have opened up as social media data has grown increasingly rich and abundant. The new data-driven methodologies based on topic models have provided academics with a new perspective on social phenomena. Because social media information is brief, text-heavy, and unstructured, methodological hurdles emerge for both data analysis and collecting. To identify which deep learning model would best categorise our data, we employ deep learning models such as Convolutional Neural Networks (CNNs), Bidirectional Encoder Representations from Transformers (BERT), and Global Vectors for Word Representation (GloVe). The data for this article comes from Mumbai Malaria posts on Twitter. In contrast to machine learning approaches, these models proved to be highly exact and beneficial in the study of emotions.

Introduction

Sentiment analysis is becoming increasingly popular among firms and researchers seeking to understand their customers' emotions. It is now simple to recognise sentiment and categorise words into positive, negative, and neutral categories utilising natural language processing, statistical analysis, and text analysis. With people sharing their opinions and reviews of businesses more openly than ever before, sentiment analysis has evolved into a valuable tool for monitoring and interpreting online interactions. You may understand what makes and disappoints your consumers by automatically analysing customer feedback and ratings based on survey replies and social media discussions.

People nowadays are increasingly linked to one another via the internet, utilising numerous platforms such as Twitter, Facebook, and Instagram. People openly discuss their views and opinions on these platforms. The sentiment analysis uses data to assist the brand understand the opinions and ideas of their customers and to make them aware of any adjustments that may need to be made to their products and services. Every day, a tremendous amount of data is created on platforms such as Twitter. The data is in numerous formats, the majority of which are unstructured. Because of the polarity, context, subjectivity, tone, emoji, irony, sarcasm, defining neutral, and comparison in the phrases, analysing sentiments can be challenging.

Topic modelling in text mining employs semantic text structures to interpret unstructured data without the usage of training data or preset tags. It mostly use unsupervised machine learning to identify clusters or groups of similar words within a text. Topic modelling is a quick and easy way to begin analysing your data that requires no training. BERTopic also enables manual and semi-supervised subject modelling, long-documents, hierarchical, class-based, dynamic, and online topic modelling, in addition to guided and supervised topic modelling. The topic modelling technique BERTopic use transformers and the c-TF-IDF to generate topics that are easy to grasp while retaining crucial phrases from the topic descriptions. In this study, we look for polarity in tweets. We employed CNNs, GloVe, and fine-tuned Google's pre-trained BERT architecture to get the best results possible for the bulk of NLP jobs.

Related Work

In this research paper[1] [Siriwat Limwattana and Santitham Prom-on] Deep Word-Topic Latent Dirichlet Allocation (DWT-LDA) is a novel method for training Latent Dirichlet Allocation utilising word embedding.They employed two datasets, Patnip and Amazon, both of which had been pre-processed.The Patnip dataset has 58304 postings and a vocabulary of 25847 terms.The Amazon dataset comprised 40000 reviews collected across the site, resulting in a final

vocabulary size of 15824 words. The result is a discrepancy between the DWT-LDA and LDA of roughly 0.28 on the Pantip dataset and 0.24 on the Amazon dataset.Their DWT-LDA technique provides more particular keywords for each subject, resulting in a clearer topic annotation.

A case study [2] Business analytics research was carried out. The system is based on topic modelling using Latent Dirichlet Allocation. They utilised it to identify various articles inside the business economy dataset. The stages are as follows: data pre-processing for cleaning and cross-validation, LDA Topic Modelling, and post-processing to construct numerous themes and their frequency of recurrence. The model is next examined in detail by computing the perplexity and coherence score between the subjects.

In this research paper [3] They suggested a co-STM, which combines collaborative text categorization with the Supervised Topic Model. This collaborative text classification system incorporates the supervised topic model SLDA (Supervised LDA). To pick credible unlabelled samples, this research employs a confidence calculation approach based on posterior probability distance and a category-based unlabelled sample selection strategy. They discovered that the co-STM text classification technique can increase semi-supervised text classification performance.

In [4] They attempted to construct an enhanced CombinedETM topic model from an existing CombinedTM topic model in their research work.To improve, they combined the TM topic model with the BERT model to extract emotional information from text. They do research and analysis on data linked to adolescents' personalities in the paper, and they also investigate the key elements that influence the development of adolescents' personalities. They developed remedies for young people's psychological difficulties based on those elements. The results suggest that the CombinedETM topic model, which incorporates emotional information, is more coherent than the basic subject model.

A study [5] Sustainable artificial intelligence for sustainable energy research was carried out. The LDA, BERT, and Clustering were then combined and assessed using contextual topic modelling. The researchers then coupled computational analyses with text analysis of linked scientific articles to determine the primary academic topics and themes. Based on the revealed theoretical gaps, 14 prospective future research threads were identified. The study employs a unique topic modelling technique to explore scientific literature and identify difficulties and potential solutions.

This study [Roman Egger and Joanne Yu] aims [6] to assess the performance of four distinct topic modelling strategies, namely Latent Dirichlet Allocation (LDA), Non-negative matrix factorization (NMF), Top2Vec, and BERTopic. They used Twitter posts as a dataset and ran several algorithms on them. The data was from November 2021, and after pre-processing, there were 31,800 tweets left. When comparing NMF and LDA outcomes, NMF emphasises judgement more than LDA.Top2Vec is strongly recommended for raw documents that will be interpreted.

The study[7] [Murimo Bethel Mutanga] intends to find out what local concerns relating to the epidemic are being discussed by individuals and what influence these issues have on regulatory compliance. The information is derived from the COVID-19-related Twitter network. For the extraction of subjects mentioned by people, the Latent Dirichlet Allocation (LDA) technique was used. Following the trial, it was discovered that users on Twitter discussed alcohol sales and consumption, remaining at home, daily statistics tracing, police brutality, 5G and vaccine conspiracy theories. As a result of bogus news circulating on social media platforms, individuals were resistant to acts that affected their income, and there was a propensity to undergo tests or immunisations.

In this study [8] different procedures of text categorization, classifying, and different techniques for extracting features in textual information are discussed. They have analyzed Bangla news comments sentiment using a hybrid approach and a pre-trained deep learning classifier. The proposed hybrid model utilizes an optimizer function “Adam” along with a word embedding “Glove”. The dataset used in the model is collected from online platform Kaggle. The dataset diverse text categorization and classification strategies, as well as diverse techniques for extracting characteristics from textual material, are described. They used a hybrid technique and a pre-trained deep learning classifier to analyse the sentiment of Bangla news comments. The suggested hybrid model employs an optimizer function called "Adam" as well as a word embedding called "Glove." The dataset for the model was obtained from the open site Kaggle. There are 13802 records in the collection. They used preprocessing techniques on the dataset to develop a model. This results in a good dataset. The proposed hybrid model combines two well-known deep learning approaches, BiLSTM and CNN. When these two techniques were compared, the suggested hybrid model outperformed the FastText model. The hybrid model achieves an accuracy of around 89.89%.

This paper [9] by [Pallab Chowdhury] about classification using various approaches. Because there was so much text available in so many various formats, the research experts acquired a lot of unstructured data. They searched the literature for alternative strategies to organise this dispersed knowledge into a predefined volume. Text mining has grown in popularity over the years. The study discusses text classification, grading, and several approaches for feature extraction in short texts, such as news categorization based on headlines. Their goal is to categorise various sorts of Bangla newspaper articles into ten distinct groups. For increased classification performance, they applied sophisticated data tokenization techniques and unsupervised 'GloVe' vectorization. Then, using the dataset, we used LSTM and CNN as our major feature extractors.. When compared to current models such as the binary SVM classifier, conventional LSTM, BiLSTM, CNN, or ANN. The attained outcome is a higher accuracy of 87%.

In this [10] study article, they presented many ways and novel approaches to overcoming this problem. To carry out the experiment, they had generated four datasets with varied degrees of imbalance. BBC News, 20NewsGroups, BBC News (Imbalanced), and Reuters-21578 (modified) are the datasets. They sought to demonstrate that the feature extraction techniques Singular Value Decomposition (SVD) and GloVe are critical to lowering the influence of disequilibrium in text categorization, particularly in ensemble and deep learning. The study found that both SVD and GloVe were important in improving the performance of neural networks and ensemble classifiers, as well as in classifying both unbalanced and balanced data.

A research [11] was conducted (Arnab Roy ,Muneendra Ojha) on Deep learning models are applied to the 2016 Twitter dataset.They had 6000 total tweets which were the training set and had a total of 2,043 tweets which were negative, 3,094 tweets which were positive, and 863 tweets which were neutral.There were 20632 tweets in the test batch. Their goal is to compare tweet sentiment categorization using deep learning models such as Google BERT, Bidirectional LSTM, and Convolutional Neural Networks (CNNs) using the SemEval-2016 dataset.The maximum accuracy they attained was 64.1% for Google BERT, which surpassed the other two models as the best pre-trained model.

In this research paper,[12] Bidirectional Long Short-Term Memory (LSTM) networks were provided, as well as a modified, full gradient version of the LSTM learning process. On the benchmark task of framewise phoneme classification, they compare Bidirectional LSTM (BLSTM) and numerous alternative network topologies. The TIMIT corpus dataset was utilised. The major conclusion is that bidirectional networks outperform unidirectional networks. Long Short-Term Memory (LSTM) is significantly quicker and more accurate than regular Recurrent Neural Nets (RNNs) and time-windowed Multilayer Perceptron's (MLPs). Their findings support the notion that contextual information is critical for speech processing and that BLSTM might be a useful tool.

In [13] Individual household electric demand forecast is far more difficult than community level electrical load prediction due to considerable uncertainty and unpredictability. The study proposes a deep learning framework based on a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM).The proposed hybrid CNN-LSTM model extracts features from input data using CNN layers and sequence learning using LSTM layers. In comparison to the LSTM-based model, the result generated an average mean absolute percentage error (MAPE) of 40.38% for individual household electric load estimates. The model was then compared to a recently constructed LSTM-based model evaluated on the same dataset, we obtained 4.01%, 4.76%, and 5.98% improvement for one, two, and six look-forward time steps, respectively.

This study aims [14] to forecast the incidence of malaria in chosen geographical regions. Satellite data and clinical data, together with a long short-term memory (LSTM) classifier, were utilised to forecast malaria abundances in the Indian state of Telangana. The suggested model provided a 12-month seasonal pattern for a few chosen state areas. Each region behaved differently depending on the local situation. The study's findings revealed that both environmental and clinical variables have a role in malaria transmission. Finally, the Apache Spark-based LSTM provides an effective strategy for locating malaria.

The main [15] The goal of strategic planning in the public health system is to minimise deaths and manage patients. The suggested forecast models for the time series prediction of confirmed cases, fatalities, and recoveries in 10 major COVID-19-affected nations are assessed in this research. Autoregressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), long shot term memory (LSTM), and bidirectional long short-term memory (Bi-LSTM) are among these models. Bi-LSTM, LSTM, GRU, SVR, and ARIMA models perform best to worst in all situations. Bi-LSTM yields the lowest MAE and RMSE values for fatalities in China, 0.0070 and 0.0077, respectively. The best r2_score value for instances retrieved in China was 0. 9997.

There[16] Many studies have been undertaken in order to develop machine-understandable representations and to use machine learning algorithms to these word representations in order for them to understand and detect the sentiment of a text. This study compares the usefulness of BERT embeddings for catastrophe prediction using Twitter data to classic context-free word embedding approaches. For the study, they obtained both quantitative and qualitative outcomes. The dataset utilised is a Kaggle dataset with two different files, totaling 7,613 tweets for training and 3,263 tweets for testing, with each row of the train data containing ID, natural language text or tweet, and label. The results show that contextual embeddings outperform standard word embeddings in catastrophe prediction challenges.

This study [by Qin Xiang Ng] [17] wanted to analyse the prevailing unfavourable views regarding COVID-19 immunisation utilising the analysis of public twitter messages over a 16-month period. From 1 April 2021 to 1 August 2022, all tweets in English were retrieved. On chosen negative sentiment tweets, the BERT (bidirectional encoder representations from transformers) was applied. After cleaning the data, a total of 4,448,314 tweets were evaluated. They discovered that six subjects were frequently associated with unfavourable attitudes towards immunisation. People's reactions to policy, safety, and efficacy were all addressed in these themes.

In this novel study [18], Using social media posts to detect influenza-related tweets may offer early warning of influenza epidemics. Deepluenza is a deep learning algorithm that can effectively detect influenza-related tweets. Tweets in both English and Arabic are utilised. They performed studies and compared the Deepfluenza results to real-world influenza data provided by health authorities. Deepluenza, which is

based on the BERT, obtained 0.99 accuracy and an F1-score of 0.98 in the influenza reporting class. The model's application demonstrated a favourable relationship between the amount of social media reports discovered and the actual number of influenza-related hospital visits. We find a link between tweets and the amount of persons who visit hospitals after utilising both.

This paper [19] wanted to learn how individuals on Twitter express their hopeful and pessimistic opinions regarding COVID-19. To extract the semantic information, they employed transformer embedding and a variety of network designs. They discovered that the best pessimistic and optimistic detection models are built on bidirectional long- and short-term memory networks. There were 150,503 tweets and 51,319 unique users. Conversations with more gloomy messages revealed just a 62.21% emotional change. In response, just 10.42% of the interactions that were more cheerful kept the atmosphere going. The emotional unpredictability of the user is also connected to social influence.

This study [by Menghan Zhang] [20] wants to determine if the sentiments of social bots influence the attitudes of COVID-19 vaccination users. The researchers then discovered social bots and created an innovative computational framework, the BERT-CNN sentiment analysis framework. To discover COVID-19 vaccination post attitudes on Twitter between December 2020 and August 2021.The goal was to investigate the effects of social bots on human vaccination attitudes online. The Granger causality test was then used to determine whether there was time-series connection between social bot emotions and human attitudes. The study discovered that social bots can impact human attitudes towards COVID-19 vaccinations. Their power to convey feelings on social media, whether favourable or bad, will have a commensurate influence on human attitudes.

This study [21] The goal of this study is to look at the BERT language model for emotion recognition in Indonesian-language Tweets. The Bert model is optimal for representing context. The strategy used is fine-tuning rather than pre-training, which takes a large amount of data and resources. Two pre-trained models were utilised to determine the model's efficacy and performance. When compared to other models, the Bert model had the highest accuracy of 77%. The benefit in this case was the minimal computing time required.

In this [22] They used a Convolutional Neural Network with Bidirectional Long-Short Term Memory (CNN-Bi-LSTM) to collect public tweets on the COVID-19 worldwide epidemic from Twitter. This hybrid Deep Learning system is used to determine if the user has good, negative, or neutral thoughts towards the epidemic. To extract word embedding, pre-processing techniques are utilised, together with a word embedding pre-trained model. The CNN-Bi-LSTM is a hybrid model that has been tested using accuracy, precision, recall, and f1 approaches. CNN-Bi-LSTM with Fast Text pre-trained model obtained 99.33% accuracy while CNN-Bi-LSTM with GloVe pre-trained model achieved 97.55% accuracy.

The aim of this study is to identify the [23] issues and attitudes in the public COVID-19 vaccine-related social media conversation. From March 11, 2020 to January 31, 2021, tweets were acquired from a large-scale COVID-19 Twitter buzz data collection. They utilised R software to clean and save the tweets. Vaccination, vaccinations, vaccines, immunisation, vaccinate, and vaccinated are the terms used. The dataset contains 1,499,421 distinct tweets from 583,499 distinct people. R is used for topic modelling, sentiment and emotion analysis, and latent Dirichlet allocation. A total of 16 subjects were planned, and vaccination progress will be more widely addressed around August 11, 2020.We are seeing an increase in favourable feeling towards COVID-19 vaccinations, and adoption of COVID-19 vaccines has increased when compared to earlier immunisations.

This study [Mohammad Mujahid] [24] Examine the impact of e-learning by assessing people's attitudes towards e-learning. They employ a Twitter dataset including 17,155 e-learning-related tweets. With the appropriateness and efficacy of machine learning models in mind, this work employs TextBlob, VADER (Valence Aware Dictionary for Sentiment Reasoning), and SentiWordNet to examine the polarity and subjectivity score of twitter text. To create and test the models, the two feature extraction approaches employed are TF-IDF (Term Frequency-Inverse Document Frequency) and BoW (Bag of Words).When applied with Bow features, the random forest and support vector machine classifiers get the greatest accuracy of 0.95. Performance is compared for TextBlob, VADER, and SentiWordNet findings, as well as classification results from machine learning models and deep learning models, such as CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), CNN-LSTM, and Bi-LSTM (Bidirectional-LSTM) are some examples. Furthermore, topic modelling is used to identify the challenges connected with e-learning, which reveals that the top three problems are uncertainty of campus opening date, children's impairments to comprehend online education, and trailing efficient networks for online education.

This study conducted by Parveen SV [25] (2022) Some examples are CNN (Convolutional Neural Network), LSTM (Long Short-Term Memory), CNN-LSTM, and Bi-LSTM (Bidirectional-LSTM). Furthermore, topic modelling is utilised to identify the hurdles associated with e-learning, revealing that the top three issues are campus opening date uncertainty, children's impairments to understand online education, and lagging efficient networks for online education.

Methodology and Implementation details

A. Data Collection

The data was gathered via Twitter API access, which is based on live tweets from Mumbai users about the rise in Malaria cases, which is a big concern for Mumbai

residents. We compiled the tweets of Mumbai users from the years 2020 to 2022. We primarily focused on positive and negative tweets on how the BMC works to prevent malaria in the city. So we gathered people's reactions to BMC works, with an emphasis on the positive and negative reactions of Mumbai twitter users to BMC.

1.Positive
2.Negative

For the sake of study the process content of Malaria Tweets of Mumbai city 2020-2022. The Data set contains three columns Text, User, and Location, here is our Dataset present below Table 1.

Text	User	Location
RT @Dharavi12: @	meee_mumbaikar	Mumbai, India
@ataulkhan09 @m	Dharavi12	Dharavi, Mumbai
No one dies from a	Ramtheunbatable	Mumbai, India
#Malaria No more	Ramtheunbatable	Thane,Mumbai
Last month, the #B	Ramtheunbatable	
#malaria case are r	Ramtheunbatable	

Table 1. EXAMPLE OF TWEETS WITH THEIR RESPECTIVE SENTIMENT

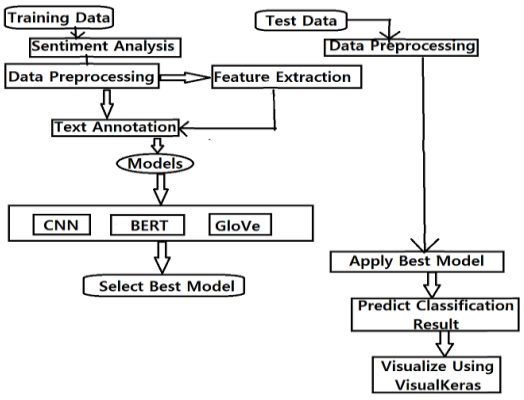

Fig1. Text Classification Methodology

1. SENTIMENT ANALYSIS

We can assess whether data is good, negative, or neutral using Sentiment Analysis. Based on sentiment analysis, it enables us to identify how individuals feel about malaria and how they react to it. Sentiment analysis is used in natural language processing (NLP) to assess the emotional tone of a text utterance. It's also known as opinion mining.

We obtained tweets from Twitter for this study using the Twitter API.

This graph depicts Twitter Sentiment Analysis. We were able to retrieve hashtags such as #Mumbai, #BMC, and #Malaria from social media networks like as Twitter by using API access. We cleaned and evaluated the neat tweets after collecting them, visualised them, and assessed the good, negative, and neutral remarks.

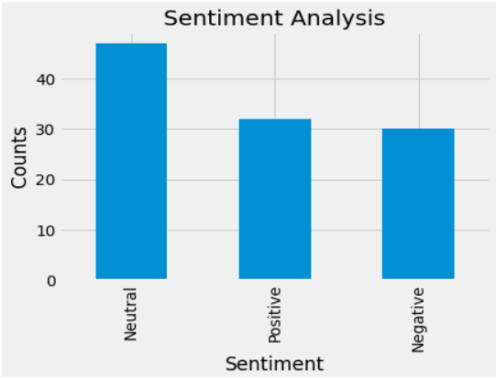

Fig 2: Sentiment Analysis

Then we plot some WordCloud Graph of each text sentiment, like Positive WordCloud, Negative WordCloud and Neutral WordCloud.

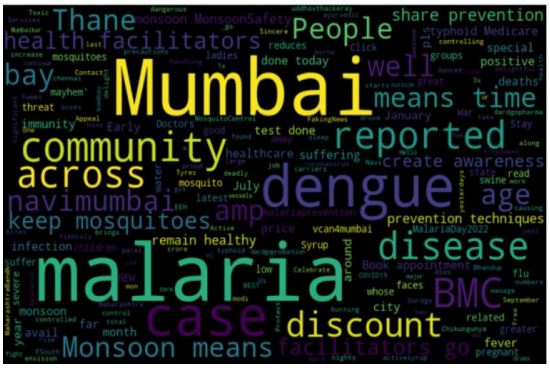

Fig 3: Positive WordCloud

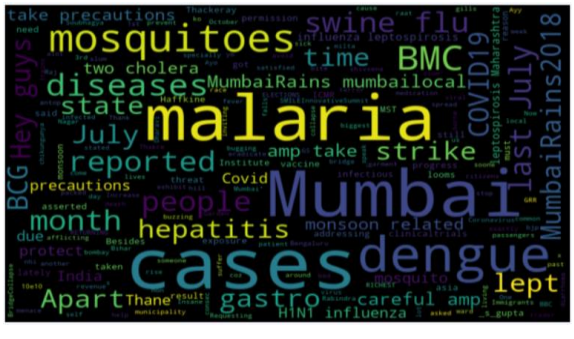

Fig 4: Negative WordCloud

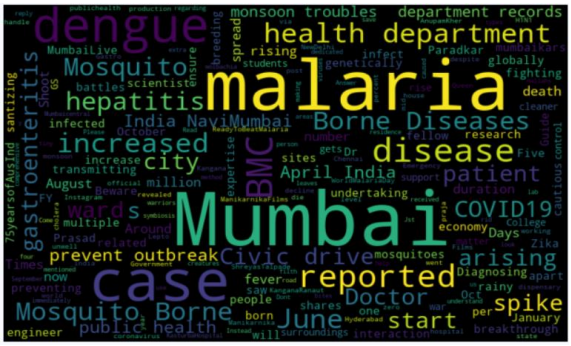

Fig 5: Neutral wordCloud

B. Classification Of Model

We built a system based on Convolutional Neural Network (CNN), BERT, and the Glove Model using the Malaria Mumbai Tweets Dataset for Sentiment Analysis. Raw tweets marked Positive, Negative, and Neutral comprise the Sentiment training set. We compared and examined the three models, and we also used VisualKeras Graph to visualise the model. We gathered the top 50 most often used terms for topic modelling by mixing tweets.

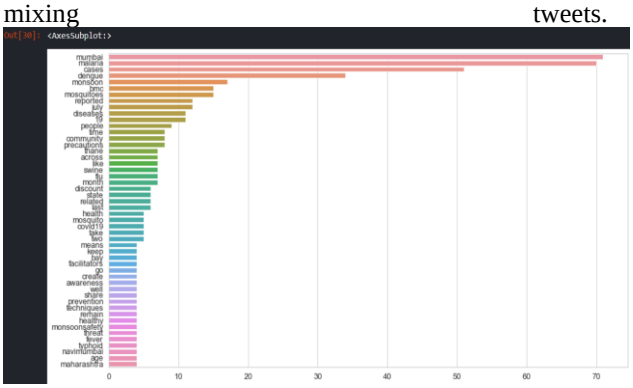

Fig. 6: Word Frequency Graph of Positive and Negative Tweets of most common 50 words

1. Convocation Neural Network (CNN)

Though CNNs are more commonly connected with computer vision difficulties, they have lately been applied in NLP with interesting results. CNNs are just many layers of convolutions with non-linear activation functions such as ReLU, tanh, or SoftMax applied to the output.

The CNN model may be recognised as the appropriate model for text data classification. It displays bigrams, trigrams, and n-grams according to the kernel size, which is the continuous sequence of words. Raw sentiment data is utilised to train the CNN model for a suitable output, which can then be trained and tested.Each kernel recognises certain phrase patterns, such as Positive and Negative, Spam and Ham emails, and so on. Most text classification tasks, like sentiment analysis, are decided by the presence or absence of particular keys somewhere in the phrase. As a result, CNN, which are strong at extracting local characteristics from data, can successfully mimic this. As a result, we picked CNN for our purpose classification assignment.

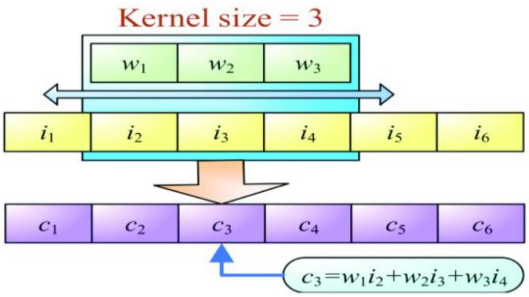

Fig.7: Convocation Neural Network (CNN)

The first step is to determine the kernel placed then calculate the matrix

w1,w2,w3 represents the size of the kernel , c3 is used to calculate the output variable and the corresponding variables are summed up with w1i2 +w2i3+w3i4.

The outputs of the self - attention layer are computed as in :

C3 = w1i2+w2i3+w3i4 …..(1)

With a batch size of 30, the model is trained for 15 epochs. The accuracy of the model is calculated and used to evaluate it. The number of right guesses divided by the total number of predictions yields the categorization accuracy.

Acc = (TP+TN)/(TP+TN+FP+FN) …..(2)

For calculating the precision, recall, f1-score:

TP/(TP+FP) ……(3)

TP/(TP+FN) ……(4)

$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

	precision	recall	f1-score	support
0	0.00	0.00	0.00	9
1	0.55	1.00	0.71	18
2	1.00	0.33	0.50	9
accuracy			0.58	36
macro avg	0.52	0.44	0.40	36
weighted avg	0.52	0.58	0.48	36

Fig .8: Classification of Convolutional Neural Network Model

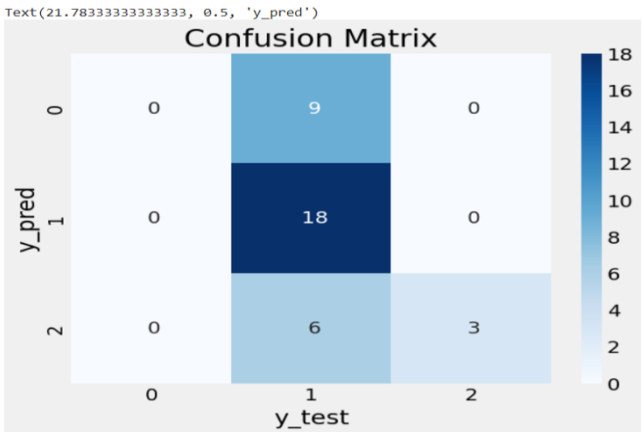

Fig. 9 Confusion Matrix of Convolutional Neural Network Model

2. Glove Model

Glove model is another term for Global Vector Model. An unsupervised learning technique generates a vector representation of words. The goal of this challenge is to detect hate speech in tweets. To make things easier, we consider a tweet to be hate speech if it incorporates racist or sexist views. Our objective in this scenario is to discern between racist and sexist messages. Essentially, given a training set of tweets and labels, we must predict the labels on the test dataset, where a label '1' signifies racism/sexism and a label '0' denotes not racism/sexism.

On the Internet, hate speech is sadly all too widespread. Social media platforms such as Facebook and Twitter frequently confront the challenge of recognising and filtering harmful messages while balancing the right to free expression. The close link between hate speech and real hate crimes emphasises the significance of recognising and controlling hate speech. Early detection of users encouraging hate speech might allow for outreach programmes aimed at preventing an escalation from speech to action. Sites like Twitter and Facebook have been aggressively combating hate speech. Despite these factors, NLP research on hate speech has been extremely restricted, owing mostly to the lack of a universal definition of hate speech, an examination of its demographic impacts, and an evaluation of the most effective aspects.

Our whole collection of tweets was divided into training and testing data in a 30:30 ratio. 30% of the testing data is public, while the remainder is confidential.The complete GloVe word embedding file must then be loaded into memory as a dictionary of word to embedding array.

Then, for each word in the training dataset, we must generate a matrix with one embedding. We may accomplish this by enumerating all unique words in the Tokenizer.word_index and retrieving the embedding weight vector from the loaded GloVe embedding.

The end result is a weighted matrix of words that we will only view during training. Following that, the model employs Bidirectional LSTM.

a) BIDIRECTIONAL LSTM\

The main principle behind bidirectional recurrent neural networks is to offer each training sequence to two independent recurrent nets, both of which are coupled to the same output layer. This means that the BRNN contains comprehensive, sequential knowledge about all points before and after it for every point in a particular sequence. Furthermore, because the net can use as much or as little of this context as needed, there is no need to define a (task-dependent) time-window or goal delay size.

b) Dropout

It is a training method in which randomly chosen neurons are disregarded. They're "dropped-out" at random. This

implies that their contribution to downstream neuron activation is eliminated temporally on the forward pass, and any weight changes are not applied to the neuron on the backward trip.

c) SpatialDropout1D

This version accomplishes the same purpose as Dropout, except instead of individual components, it dumps whole 1D feature maps.

Next to calculate the Accuracy and Confusion Matrix of GloVe Embedding Model

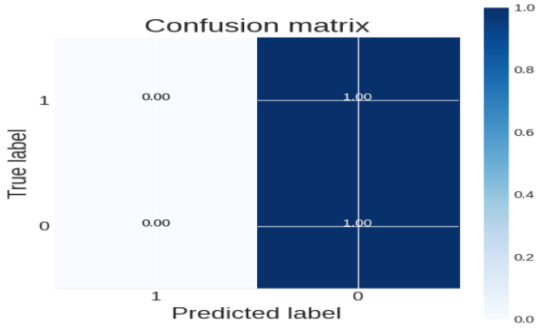

Fig.10: Confusion matrices of Glove model

	precision	recall	f1-score	support
0	0.00	0.00	0.00	9
1	0.44	1.00	0.61	7
accuracy			0.44	16
macro avg	0.22	0.50	0.30	16
weighted avg	0.19	0.44	0.27	16

Fig. 11: Classification of GloVe model

3. BERT

During 2018, Jacob Devlin and his Google colleagues created BERT, a sophisticated machine learning model based on Transformers. For many natural language applications, BERT is a good pre-trained language model that enables computers to learn effective representations of text when used in context, outperforming the current state of the art. BERT is an acronym that stands for Bidirectional Encoder Representations from Transformers. In the BERT design, each Transformer encoder is made up of a feed-forward and a self-attention layer that are layered on top of each other. After merging Positive and Negative tweets and initialising with 1 and 0, our model was trained using the Malaria Mumbai Tweets dataset. After the pre-processing is complete, the BERT pre-processor and encoder must be downloaded in order to produce the model. For our model, which consists of one dense layer and one output unit, the sigmoid function is employed as the output unit. We can attain an accuracy of 63.0% using the training dataset. Because of the volatility of the training process, the accuracy might vary somewhat from one epoch to the next.

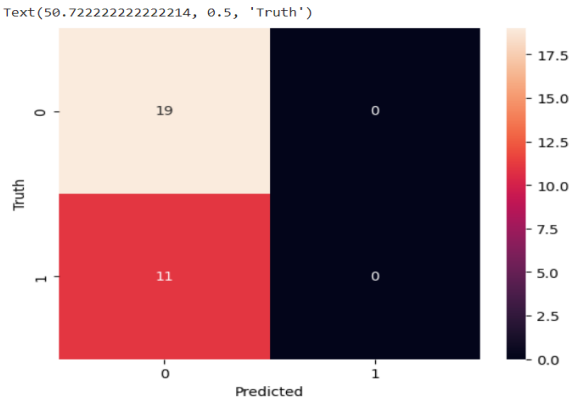

Fig 12. Confusion Matrices of BERT Model

	precision	recall	f1-score	support
0	0.63	1.00	0.78	19
1	0.00	0.00	0.00	11
accuracy			0.63	30
macro avg	0.32	0.50	0.39	30
weighted avg	0.40	0.63	0.49	30

Fig 13. Classification Report of BERT Model

Experimentation setup and Results

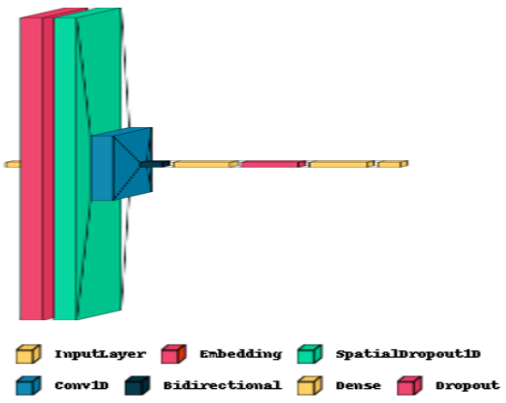

Fig.14: CNN VisualKeras Model

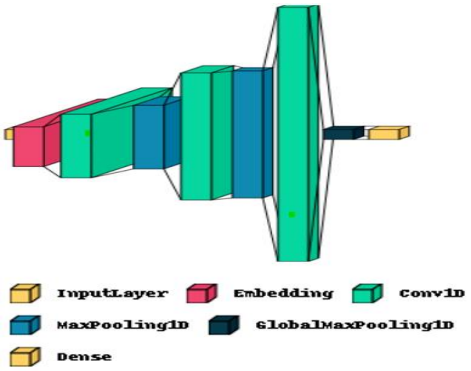

Fig.15: GloVe VisualKeras Model

When three models are compared, the BERT Model is determined to be the best match, with a high accuracy of 0.63%. Although we cannot claim that 0.63% accuracy is the greatest, it is greater than other models such as CNN and GloveEmbedding. Machine learning and deep learning models can assist make accurate predictions, but they are less valuable unless interpretability is included in the process. Deep learning models become more intelligible when they are visualised. We will visualise deep learning models using visualkeras, a Python visualisation library. It allows us to visualise deep learning models built using the Keras API, as well as their network topologies built with Tensorflow and Keras.

Conclusion

Our study used three well-known deep learning models for Twitter sentiment analysis, and we pre-processed the data to improve model accuracy by reducing noise. After comparing and assessing the models, we discovered that the BERT model outperformed the others. Because the BERT model considers and combines all words in the sequence, it is strong because it gives a deeper knowledge of context rather than merely predicting the next word.

Future Scope

In the future, we will use many deep learning models to try to increase the model's accuracy. And experiment with several model datasets on a live dataset. In addition, we will utilise tweet location as a Geo Location and build a model based on the information.

References

[1] S. Limwattana and S. Prom-on, "Topic Modeling Enhancement using Word Embeddings," 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE), Lampang, Thailand, 2021, pp. 1-5, doi: 10.1109/JCSSE53117.2021.9493816.

[2] V. Vukanti and A. Jose, "Business Analytics: A case-study approach using LDA topic modelling," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1818-1823, doi: 10.1109/ICCMC51019.2021.9418344.

[3] G. Zhang, H. Zheng and X. Liu, "co-STM text categorization method based on Supervised Topic Model," 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Changsha, China, 2021, pp. 462-467, doi: 10.1109/AEMCSE51986.2021.00101.

[4] C. Chen, Z. Chen, T. Hu, J. Ge, H. Peng and S. Liu, "An Integrating Emotional Information and Topic Model Method for Text Topic Mining," 2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom, 2021, pp. 450-454, doi: 10.1109/NetCIT54147.2021.00095.

[5] Tahereh Saheb, Mohamad Dehghani, Tayebeh Saheb, Artificial intelligence for sustainable energy: A contextual topic modeling and content analysis, Sustainable Computing: Informatics and Systems, Volume 35, 2022, 100699,ISSN 2210-5379,https://doi.org/10.1016/j.suscom.2022.100699

[6]Egger R, Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Front Sociol. 2022 May 6;7:886498. doi: 10.3389/fsoc.2022.886498. PMID: 35602001; PMCID: PMC9120935.

[7] Murimo Bethel Mutanga, Abdultaofeek Abayomi "Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach "Journal Article, 2022, https://journals.co.za/doi/abs/10.1080/20421338.2020.1817262

[8] U. Saha, M. S. Mahmud, A. Chakrobortty, M. T. Akter, M. R. Islam and A. A. Marouf, "Sentiment Classification in Bengali News Comments using a hybrid approach with Glove," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 01-08, doi: 10.1109/ICOEI53556.2022.9777096.

[9] Chowdhury, P., Eumi, E.M., Sarkar, O., Ahamed, M.F. (2022). Bangla News Classification Using GloVe Vectorization, LSTM, and CNN. In: Arefin, M.S., Kaiser, M.S., Bandyopadhyay, A., Ahad, M.A.R., Ray, K. (eds) Proceedings of the International Conference on Big Data, IoT, and Machine Learning. Lecture Notes on Data Engineering and Communications Technologies, vol 95. Springer, Singapore. https://doi.org/10.1007/978-981-16-6636-0_54

[10] T. Hossain, H. Zahin Mauni, and R. Rab, "Reducing the Effect of Imbalance in Text Classification Using SVD and GloVe with Ensemble and Deep Learning", *Comput. Inform.*, vol. 41, no. 1, pp. 98–115, Apr. 2022.

[11] A. Roy and M. Ojha, "Twitter sentiment analysis using deep learning models," 2020 IEEE 17th India Council

International Conference (INDICON), New Delhi, India, 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342279.

[12] Alex Graves, Jürgen Schmidhuber,"Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks",Volume 18, Issues 5–6,2005,Pages 602-610,ISSN 0893-6080,https://doi.org/10.1016/j.neunet.2005.06.042.

[13] M. Alhussein, K. Aurangzeb and S. I. Haider, "Hybrid CNN-LSTM Model for Short-Term Individual Household Load Forecasting," in IEEE Access, vol. 8, pp. 180544-180557, 2020, doi: 10.1109/ACCESS.2020.3028281.

[14] Thakur Santosh, Dharavath Ramesh, Damodar Reddy,LSTM based prediction of malaria abundances using big data,Computers in Biology and Medicine, Volume 124,2020,103859,ISSN 00104825,https://doi.org/10.1016/j.compbiomed.2020.103859.

[15] Farah Shahid, Aneela Zameer, Muhammad Muneeb,Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM,Chaos, Solitons & Fractals,Volume 140,2020,110212,ISSN 09600779,https://doi.org/10.1016/j.chaos.2020.110212.

[16] Sumona Deb, Ashis Kumar Chanda,Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data,Machine Learning with Applications,Volume 7,2022,100253,ISSN 2666-8270,https://doi.org/10.1016/j.mlwa.2022.100253.

[17] Q. X. Ng, S. R. Lim, C. E. Yau, and T. M. Liew, "Examining the Prevailing Negative Sentiments Related to COVID-19 Vaccination: Unsupervised Deep Learning of Twitter Posts over a 16 Month Period," *Vaccines*, vol. 10, no. 9, p. 1457, Sep. 2022, doi: 10.3390/vaccines10091457. [Online].

[18] Balsam Alkouz, Zaher Al Aghbari, Mohammed Ali Al-Garadi, Abeed Sarker,Deepluenza: Deep learning for influenza detection from Twitter,Expert Systems with Applications,

Volume 198,2022,116845,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2022.116845.

[19] Guillermo Blanco, Anália Lourenço,Optimism and pessimism analysis using deep learning on COVID-19 related twitter conversations,Information Processing & Management,Volume 59, Issue 3,2022,102918,ISSN 0306-4573,https://doi.org/10.1016/j.ipm.2022.102918.

[20] M. Zhang, Z. Chen, X. Qi, and J. Liu, "Could Social Bots' Sentiment Engagement Shape Humans' Sentiment on COVID-19 Vaccine Discussion on Twitter?," *Sustainability*, vol. 14, no. 9, p. 5566, May 2022, doi: 10.3390/su14095566. [Online]. Available: http://dx.doi.org/10.3390/su14095566

[21] K. S. Nugroho and F. A. Bachtiar, "Text-Based Emotion Recognition in Indonesian Tweet using BERT," 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2021, pp. 570-574, doi: 10.1109/ISRITI54043.2021.9702838.

[22] T. T. Mengistie and D. Kumar, "Deep Learning Based Sentiment Analysis On COVID-19 Public Reviews," 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Korea (South), 2021, pp. 444-449, doi: 10.1109/ICAIIC51459.2021.9415191.

[23] Lyu J, Han E, Luli G,COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis,J Med Internet Res 2021;23(6):e24435,URL:https://www.jmir.org/2021/6/e24435,DOI: 10.2196/24435

[24] M. Mujahid *et al.*, "Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19," *Applied Sciences*, vol. 11, no. 18, p. 8438, Sep. 2021, doi: 10.3390/app11188438. [Online]. Available: http://dx.doi.org/10.3390/app11188438

[25] P. SV *et al.*, "Twitter-Based Sentiment Analysis and Topic Modeling of Social Media Posts Using Natural Language Processing, to Understand People's Perspectives Regarding COVID-19 Booster Vaccine Shots in India: Crucial to Expanding Vaccination Coverage," *Vaccines*, vol. 10, no. 11, p. 1929, Nov. 2022, doi: 10.3390/vaccines10111929. [Online]. Available: http://dx.doi.org/10.3390/vaccines10111929