New Generation Computing (2022) 40:1029–1052
https://doi.org/10.1007/s00354-021-00151-1

NEW GENERATION COMPUTING

A Transformer-Based Model for Evaluation of Information Relevance in Online Social-Media: A Case Study of Covid-19 Media Posts

Utkarsh Sharma[1] · **Prateek Pandey**[1] · **Shishir Kumar**[2]

Received: 17 September 2021 / Accepted: 26 December 2021 / Published online: 10 January 2022

Abstract
Online social media has become a major source of information gathering for a huge section of society. As the amount of information flows in online social media is enormous but on the other hand, the fact-checking sources are limited. This shortfall of fact-checking gives birth to the problem of misinformation and disinformation in the case of the truthfulness of facts on online social media which can have serious effects on the wellbeing of society. This problem of misconception becomes more rapid and critical when some events like the recent outbreak of Covid-19 happen when there is no or very little information is available anywhere. In this scenario, the identification of the content available online which is mostly propagated from person to person and not by any governing authority is very needed at the hour. To solve this problem, the information available online should be verified properly before being conceived by any individual. We propose a scheme to classify the online social media posts (Tweets) with the help of the BERT (Bidirectional Encoder Representations from Transformers)-based model. Also, we compared the performance of the proposed approach with the other machine learning techniques and other State of the art techniques available. The proposed model not only classifies the tweets as relevant or irrelevant, but also creates a set of topics by which one can identify a text as relevant or irrelevant to his/her need just by just matching the keywords of the topic. To accomplish this task, after the classification of the tweets, we apply a possible topic modelling approach based on latent semantic analysis and latent Dirichlet allocation methods to identify which of the topics are mostly propagated as false information.

Keywords Covid-19 · Tweet classification · Text mining · Online social media · BERT

✉ Utkarsh Sharma
utkarsh_shar@yahoo.co.in

Extended author information available on the last page of the article

Introduction

On July 1, 2018, five men were brutally murdered by a raging mob in Rainpada village of Maharashtra's Dhule district, in India. Their death was triggered by the rumours related to men kidnapping children and selling their body parts for money circulated on the messaging platform WhatsApp. Although those five men were not related to any of such crimes as later identified in police investigation also the gory images shown in the messages circulated on WhatsApp were from Syria where some kids were killed during a chemical attack in 2013 [1]. A tragic end of the lives of those five men just because of the propagation of misinformation on online social networks (OSN).

This was not the first incident of loss of lives due to rumour-based mob lynching, there has been a steep rise in the cases of such crimes in the recent past. But the major concern that comes to mind that what is the main cause of such a huge increase in these false information related incidents recently. People are using online social media for a very long time dated back to the year 1990 [2], but these incidents of false news propagation escalated in early 2000 when the number of users generating data on online social networking sites such as Facebook, Instagram, etc., increases to millions and close to billion also. If we talk about the monthly active users (MAU) of these platforms, then Only Facebook has more than 1.5 billion MAU along with YouTube which also has more than 1 billion MAU [3]. As the OSN market grows to many folds in recent years so as the power to generate or disseminate information goes into the hand of almost every individual on this planet. Now any individual can post or share any information whether relevant or not whether true or false without giving any thought process to identify that information is true or not by just simply clicking one button. Almost every social media platform has this functionality of resharing one another's post which makes this problem more difficult to identify that what was the exact source of this false information.

Generally, this situation is considered to be the problem of fake news, but we cannot consider this type of false information as fake news. Although fake news is not a new concept its root belongs to thirteenth century BC [4] and fake news is defined as something which is either false or contains something which is intentionally misleading the information in some other context [5]. One of the major problems with identifying the truth about articles is that often some part of the post is true and some part is fake, this type of condition arises when someone unintentionally spread the wrong news just to spread awareness but end up spreading the fake news due to shortage of evidence [6]. Sharing of fake news or irrelevant content on OSN not only brings a state of confusion in the minds of the user who reads it but it also affects the several decision-making systems built based on OSN content such as the location mapping system developed by Fan et al. [7] based on tweets collected for a natural disaster. Fake News in social media has some serious impact on share market returns also, Cepoi [8] demonstrated an asymmetric relationship between news circulating about Covid-19 and the share market gains with different impacts on high, medium and small-scale shares.

The Shorenstein Centre of the Harvard Kennedy School characterizes fake news as "deception that has the features of conventional news media, with the assumed related publication measures," perceiving the requirement for the improvement of terminology to help researchers study this method [9]. The sharing of fake news on social media is also influenced by the factor of speed by which a piece of new information disseminates about any topic, as there is a lack of information available to verify the validity or truth about the news [10], as in the case of pandemic situations like Covid-19. In March 2020 when the global pandemic of coronavirus was declared by the World Health Organization. Only in February 2020, the Director-General of WHO said that "we are not just fighting an epidemic, we are fighting an infodemic". As the condition of Covid-19 becomes worse than a normal virus spreading among continents it was declared Pandemic by WHO, along with this the fake news dissemination became a big concern to handle and it was also considered to be an "infodemic" which has to be handled with utmost priority as we handle the pandemic. The reason that this fake news war has to be prevented is that nowadays this digital media is reaching into the hands of people who do not have enough measures to clarify or judge the content that they see on social media and they immediately end up believing that content which may or may not land them in a problematic situation [11].

This is not the first time that we are dealing with a deadly virus spreading among several countries, previously we dealt with SARS which was a newly discovered human disease in early 2002 with initial occurrences in Southern China. The causing agent for SARS was identified as SARS-CoV an unrecognised strain of coronavirus in record time [12]. But the spread and harmful effects of fake news were evident at an enormous level this time when the entire population of the world was struggling with the wrath of the Covid-19 pandemic. Unfortunately, social media use (especially problematic social media use) may give rise to psychological distress. Talks about the fear of COvid-19 and its consequent effects on the mental health of people such as anxiety and insomnia. Several rumours surfaced in this duration as people died of drinking access alcohol because of a misconception that alcohol can kill the corona virus [13].

For understanding the concept of false information, we should clearly understand the difference between the terminological difference between misinformation and disinformation. Misinformation is the phenomenon when something is conceived as wrong because of their inappropriate context or some false connection of information while on the other hand disinformation is considered to be the phase where someone knowingly spread fake news to harm or get some benefit of it. The reason that the study of fake content on online social media will be more prominent in the coming future is because of the trust issue people have with the news content available mostly believed to be paid to benefit some political agenda. It is not the case that political influenced post does not exist on online social media but still there is a lot of content available that comes directly from the vision of a common man about the problems or issues faced by a common man [14].

The classification of a text as fake or real is not a straightforward approach, and things become more complex when we talk about the text available on social media platforms because of the heterogeneity of categories of users generating

the text. Generally, not all users provide facts on their social media accounts, the text may be intended to generate humour or that may be a personal view of any individual on some topic. The verdict is that not every post available online can be considered as a valid candidate to be categorized as true or false. We cannot state a text as relevant or irrelevant just based on the presence of certain words, the main thing to consider is the context of the word present in the sentence. The same words can present different meanings in different contexts. The transformer-based models like BERT are capable of mapping the context to the words and then applying the ranking to the words. The proposed model is thus first classifying the tweets as relevant or non-relevant, this classification identifies those tweets as relevant which contain some information that can be either stated as true or false and all other tweets into the non-relevant class which can never be considered as facts. For this purpose, we use the BERT (Bidirectional Encoder Representations from Transformers) classification model which is proved to be best in performance in terms of classification accuracy and for contextual mapping. For data collection, we used the Twitter API which is free to use for collecting the tweets daily. We started our data collection process from the early days of this outbreak of a pandemic, we collected a total of 40,000 tweets from 15th March 2020 to 30th July 2020 for the related hashtags like #Covid, #Corona, #Coronavirus, #Covid-19. Then we labelled the tweets based on their relevance to the topic of discussion as Relevant and Irrelevant tweets, an example of such labelled tweets is shown in Table 1. Text-based classification models suffer from the problem of heterogeneity in the text and large veracity of words and sentence pairs which makes it difficult for any text classification model to remain accurate in the long run of time or we need to keep training the model at a periodic interval of time. To somehow reduce this limitation of the models, we propose a filtering of topics also which tends to be perceived often as irrelevant as compared to the others. This can help in the filtering of social media posts in future just based on their topic belonging or providing a suitable warning with such posts. For the topic modelling task, we used the Latent semantic analysis and latent Dirichlet allocation methods and compared their performances also. Both of these algorithms provide a good measure of the frequently used word in the entire text based on which the topics are determined. A graphical representation of both topic modelling approaches is also presented in the result section.

Table 1 Sample of labelled tweets for model training

Tweet	Label
@ramainoane @nthakoana happiness is in the air & corona non-existent	Irrelevant
Guy yesterday walks into pub in Hackney: pint of corona I will get the virus later	Irrelevant
Lockdown is work, we are supporting your decision, but stop train because corona viruses is raining no a few monthsâ€¦ https://t.co/2k4A5hiJB9	Relevant
National Guard chief tested negative for corona today after an initial positive test result ðŸ‘‡ https://t.co/HYoxVdIzpj	Relevant

Related Work

Several studies have been performed related to the effects of social media on the mental, social and economic health of people during the events of such pandemic. Apparently, this is not the first time that we are dealing with a deadly virus spreading among several countries, previously we dealt with SARS which was a newly discovered human disease in early 2002 with initial occurrences in Southern China. The causing agent for SARS was identified as SARS-CoV an unrecognised strain of coronavirus in record time [12]. Similarly, there was a history of a deadly virus outbreak in different parts of the world like the ZIKA virus outbreak in 2007 [15], EBOLA virus outbreak in 2014–16 [16], MERS-COV in 2012 and many more [17]. Along with these pandemic outbreaks, there have been studies around these events such as Shin, SY et al. [18] presented a correlation study on MERS (Middle East respiratory syndrome) spread with Twitter trends and Google search reports was conducted and it was found that a high correlation of about 70% was discovered in between the search keywords and the rise in several cases of MERS in Korea. Although the study was based on only initial trends and the startling rise in cases of disease therefore it does not demonstrate both rise and fall of the trend.

A study about the spread of fake news is performed in [19] for the articles available online related to the Zika Virus outbreak in America during 2015–16. The articles were categorized mainly into three classes which are verified, rumour and satire. A topic-wise study is also performed to identify which topic has more share in fake news articles. The study shows that the stake of fake news articles is more as compared to the verified articles. In the article presented by Kaiyuan Sun et al. [20], the authors argued about the effects of early availability of epidemiological data in case of controlling the outbreak of pandemics such as Covid-19, Ebola, SARS, etc. The study was focused on cases of Covid-19 in Mainland China for January 2020. A relative risk factor is calculated to show the severity of this disease in different age groups. As this Covid-19 pandemic span over the entire world for more than a year the number of studies performed on this virus is particularly high as compared to the previous virus spreads. The spreading of fake news during such a stressed period can bring serious mental issues in the public evident from the studies performed such as Wang et al. presented an early day's study on the psychological impact of the Covid-19 pandemic on the people of China, the study was based on survey and sampling in which more than half of their subjects experienced a moderate-to-severe level of psychological impact on their lives and about one-third of their subject responded with an anxiety level of moderate-to-severe [21]. Guntuku et al. [22] shows that there was a rise in anxiety level, feeling of loneliness and stress level during the pandemic period in the people of America based on the study conducted on the tweets collected. Also, a growth in negative sentiment was visible in the tweets for that period. Bastick [23] shows that exposure to fake news articles can usurp the mental behaviour of people based on the study performed on 223 students and recording their responses pre-test and post-test. Further, we represented a tabular comparison in Table 2 of some of the literature consisting of work related to the identification of relevance of information in case of some disastrous event.

Table 2 Literature summary of previous work on the identification of relevant information on social media

Author	Technique	Dataset	Key finding	Limitations
Kwon et al. [24]	Tweets related to social distancing were partitioned into six facets depending on their applicability to society	Tweets related to the hashtag "coronavirus", collected for 3 months from January to March	Social distancing is described as six facets and their spatio-temporal analysis to the different states of the US is represented	Only a single keyword is used for crawling tweets (#coronavirus)
Singh et al. [25]	A sentiment classification model is developed based on a BERT classifier with average likes over the period, average retweets over the period, intensity analysis, polarity and subjectivity, and word cloud as the five metrics for classification	Tweets scraped for 80 days from Jan 2020 to April 2020 with hashtags #COVID2019 OR #COVID19 OR coronavirus maintained as two separate datasets one for the entire world and second for India	The model achieved an accuracy of 94% with overall neutral sentiments worldwide and a few negative sentiments from the dataset of India	Classification based on only sentiment analysis. Tweets were collected in a very early period of the pandemic when the spread was not worldwide
Pinto et al. [26]	A model has been developed to identify the relevance of the posts in social media based on the expert based initial tagging of the dataset into 7 different labels. Then text-based classification algorithms have been applied to carry on the prediction	A total of 941 documents comprising of social media posts from Twitter, Facebook posts, Facebook comments tagged as a relevant, interesting, controversial, meaningful, novel, reliable and wide scope	Classification is done on three different parameters (1) with linguistic features (65% accuracy with Naïve Bayes or SVM), (2) based on the initial prediction of six journalistic criteria (79% accuracy of Random forest)	Only traditional classification algorithms were used in the study for the classification
Rudra et al. [27]	A summarization model is developed to categorize the tweets related to an epidemic like Ebola or MERS. Tweets will be categorized as disease-related-symptoms, prevention, disease transmission, treatment, death report and non-disease tweets	Collected 200,000 Tweets each related to Ebola and MERS using the AIDR Platform	The proposed model can achieve 80% accuracy for in-class classification and 75% accuracy for the cross-domain scenario	The size of the training dataset was only 2000 tweets

Table 2 (continued)

Author	Technique	Dataset	Key finding	Limitations
Madichetty [28]	Classification of informative tweets during disasters has been done by using CNN based feature extraction and ANN-based classifier	Tweets about the natural disaster Hurricane Harvey are collected from August 26, 2017, to September 20, 2017, with an 80:20 ratio used for training and testing	The proposed method achieves an accuracy of 75.9 over the existing machine learning approaches that use unigram, bigram and trigram features	The dataset was collected for a very short period of time (less than a month)
Bhoi et al. [29]	An LSTM + CNN based classification model is used to identify the relevant tweets in case of some disaster along with a ranking of tweets and also a mapping regarding some essential service requirements stated in the tweets	FIRE-2016 dataset of 49,913 tweets regarding Nepal earthquake 2015 is used and 43,816 non-disaster-Tweets are crawled from the Twitter free API	The proposed model attains an accuracy of 89.47 which as compared to other approaches found to be the highest	Training dataset classified based on the author's perceived classes rather than using linguistic-based criteria

Proposed Method

This section describes the method used in the proposed model; the entire process is primarily divided into three modules—(1) dataset creation, (2) classification of Tweets, (3) identification of topics. We now discuss these modules in detail for further explanation.

Dataset Creation

The Twitter platform was selected to carry out the analysis of social media content in this work due to its high availability of content and also there huge popularity of Twitter among various cadres of the society. The Twitter APIs are available with different features and functionalities such as standard, premium and Enterprise levels. For our work, we used the standard Twitter API which is easy to use and also freely available without any cost (with a limitation of the number of tweets crawled per day). We crawled the 40,000 tweets from 15th March 2020 to 30th July 2020 which was considered as the first wave of Covid-19 in India, we focussed on the tweets which contain the related hashtags like #Covid, #Corona, #Coronavirus, #Covid-19. The language we considered was English for keeping the model simple. Then the next task was to label the tweets as relevant or irrelevant based on the content of the tweet. But as the task seems pretty easy it is not that simple to identify the relevance of some text because the definition of relevance may change from person to person. The notion of relevance is highly subjective and also situational so some posts might be relevant to one person but they might not be relevant to some other person. To address this problem of identifying the relevance of, we followed the scheme proposed by Pinto et al. [26], they described six different criteria based on journalism for measuring the relevance of any text. The criteria are defined as follows:

- Interestingness: does the tweet finds the attention of the audience or not.
- Controversy: does the tweet might raise some issues leading to raged discussion.
- Meaningfulness: does the tweet is valuable or not.
- Scope: does the tweet mean for a common audience or a particular group of people.
- Reliability: does the information in the tweet sounds credible.
- Novelty: does it contain some new fact or not.

Each criterion is then assigned a score of $+1$ or -1 based on the usability of that tweet and then a final score is calculated for the tweet by summing up all the scores of six factors. After the assessment, a positive score (≥ 0) is considered to be a relevant text while a text with a negative score is considered to be irrelevant. For gaining more heterogeneity of opinions, the help of five student volunteers was taken to score the tweets, then the median score out of those five evaluations was used to identify the final score. A sample of the labelled tweets is shown in Table 1. After the final labelling is done, the next step was to pre-process the tweets before

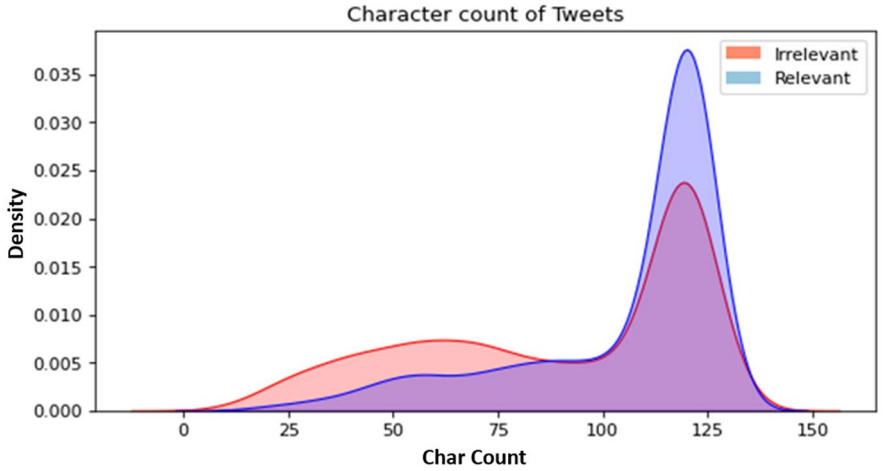

Fig. 1 Character count comparison of labelled tweets

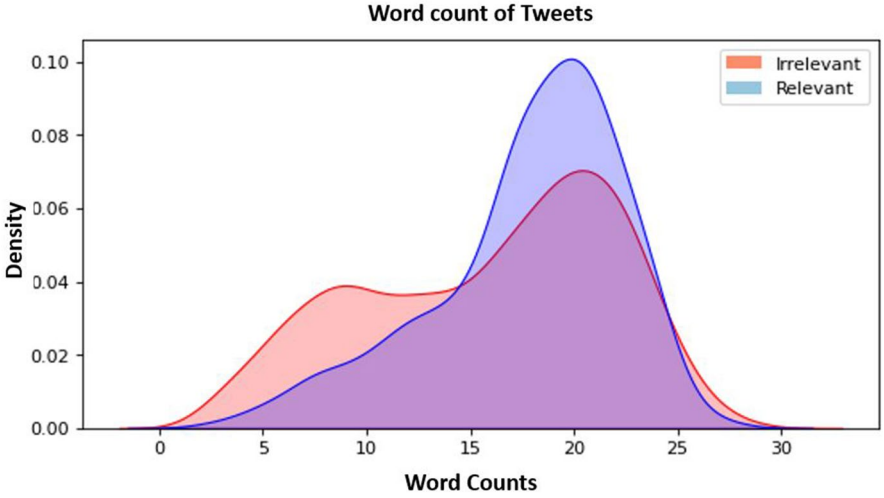

Fig. 2 Word count comparison of labelled tweets

feeding them for the processing which includes removing stop words, URLs, Mentions, Hashtags, Special characters, emojis, etc. Also, users most often retweet some tweets to remove the duplicity in the database we removed the retweets also. Before classification, the dataset has to be balanced for appropriate results and as the ratio of irrelevant tweets is more as compared to the relevant tweets, we had to apply balancing to the dataset. For balancing the dataset, we used the SMOTE [30] balancing algorithm which is based on data augmentation for the minority class. A visualisation of character count and word count of both the relevant and irrelevant tweets is shown in Figs. 1 and 2.

Word Embedding

An essential step for any text processing system is to extract the features from the text and to convert the text into a set of vectors so that the machine can understand the text and can distinguish or provide different weightage to different text. There are several methods by which one can convert the input text into a set of vectors such as Word2vec, BERT embedding, Glove, Tf-IDF n-gram, etc. [31], some modifications of these basic techniques are also suggested in the literature [32]. The embedding methods which we used in our work are namely Tf-IDF (unigram, bigram, trigram), Word2vec model and Bert embedding. The Tf-IDF method is a simple way of providing weight to different words in a text. Where "Tf" stands for term frequency and "IDF" stands for inverse document frequency and the combination of both provides a measure of the importance of a word inside the entire corpus. It provides a good score to a word that appears not too frequently in the document and is considered to be important by the model. This method can be extended using the n-gram model where n stands for the number of words in pair. So, a unigram model is a simple Tf-IDF model where a bigram model computes the importance of different two-word pairs in the text corpus and similarly a trigram model computes the scores for every pair of three-word in the text corpus. The next method word2vec [33] was developed to overcome the shortcomings of the standard Bag-of-word model by combining the techniques of continuous Bag-of-word and skip-gram. This model understands the linguistic structure of the sentence and words with similar meaning are placed together and words with dissimilar meaning are placed far apart as vectors. The major benefit of this model is that it not only depends on the collection of different words available in the statement but it understands the context of the sentence and creates the vector on basis of that only.

Another method of embedding that provided the best result in our work was BERT (Bidirectional Encoder Representations from Transformers) embedding proposed by Devlin et al. [34]. BERT is based on the transformer model of finding out the contextual relationship between the words of a sentence. The understanding of the context of the sentence comes from the bidirectional nature of the transformer which is in contrast to its previous models which processed text in either left–right or right–left directions. It also uses the concept of the masked language model, which is a text modelling technique to train the machine by the surroundings of the words. A detailed description of the BERT model is given in the next sub-section also.

BERT Model

BERT is a high-level model used for word embedding dependent on the architecture of the pre-trained transformer encoded model. We use BERT as a sentence encoder, which can precisely get the context portrayal of a sentence also we used BERT for the classification of tweets using the next sentence prediction module (NSP). BERT uses a Mask language model (MLM) to eliminate the unidirectional training

limitations and make it possible to learn from the surroundings. It arbitrarily masks a portion of the tokens from the input text furthermore, identifies the token id based on the context of the text. MLM has expanded the ability of BERT to beats when contrasted with past techniques for word embedding. It is a bidirectional framework that is equipped for dealing with the unlabelled content by mutually conditioning on both right and left context present in every layer. A deep bidirectional model is computationally more effective than a shallow right-to-left and left-to-right model. BERT supports two types of models for computation, i.e., the BERT-base model and the BERT-large model. Both of these two models were comprised of three basic components that are transformer block (L), hidden size (H), and self-attention heads (A). The values of these parameters for both the BERT-base and BERT-large model are defined in Table 3.

The processing of the BERT model is divided into two separate tasks:—(1) BERT pre-training and (2) BERT fine-tuning. The pre-training of BERT (Fig. 3) is what makes it different from the traditional unidirectional models. Here the training occurs in both directions to better understand the context of the sentence, for pre-training BERT makes use of the mask language model (MLM) and next sentence prediction (NSP). In the MLM task, some percentage of the input is masked and the model tries to predict that part. The input to the BERT pre-training phase needs to be modified into tokens and there are some special tokens also present such as *CLS*—this token denotes the starting of the sentence, *SEP*—this is the separator token used for separating two different sentences, if there is a single sentence present then it is normally appended to the last, *MASK*—this token denotes the word which is masked. In the NSP phase, the model is trained to predict the next sentence based on the binarized pre-training, this phase is much useful for the tasks of question answering system and natural language inference.

The fine-tuning phase uses the transformer feature of self-attention to tune the model by interchanging the input and output pairs and adjusting the parameters. By doing so, it performs cross attention between two sentences in a bidirectional manner as it encodes the text pair which is concatenated along with self-attention. The BERT model used in our work for classification is shown in Fig. 4.

Topic Modelling

The classification of tweets using text-based features is a good approach to filter the social media posts containing relevant text or not but to apply these classification models to the existing online social media posts will be a very difficult

Table 3 Parameters for BERT-base and BERT-large models

	BERT-base	BERT-large
Transformer block	12	24
Hidden layers	768	1024
Self-attention heads	12	16
Output parameters	110 M	340 M

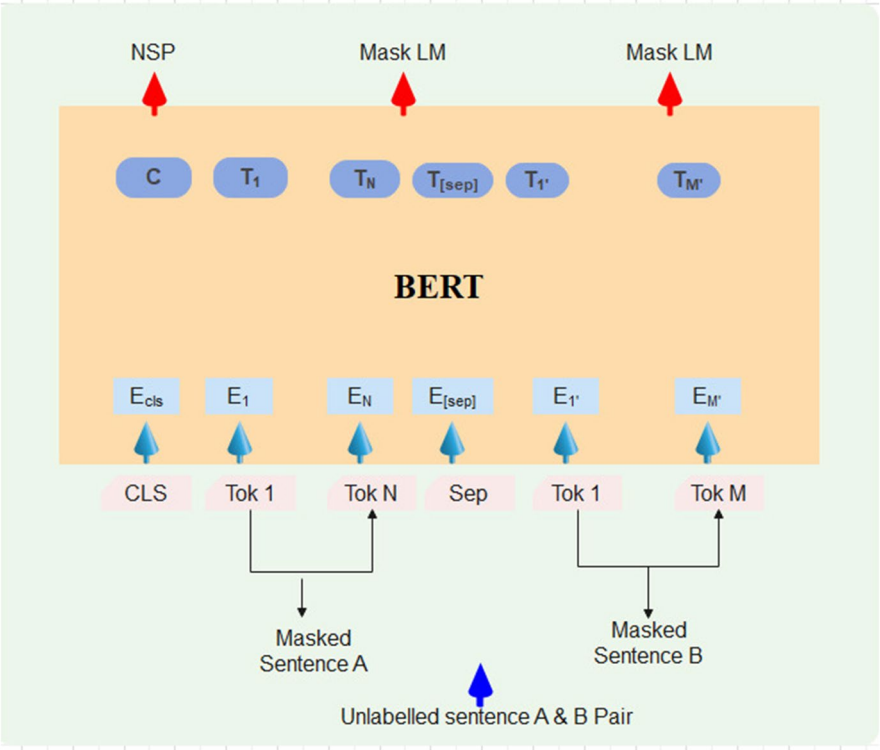

Fig. 3 BERT pre-training module for unlabelled sentence pairs

and expensive task as these models depend on the textual features and as the text may change from person to person or over time these features might not appear crucial to the classification. To provide a mechanism that can solve the problem of identification of relevancy and also supports an approach that can be easily applied to the existing models of online social media platforms, we used the concept of topic modelling. Topic modelling is the approach that can extract some common word groups from the text, which can be considered as topics to categorize the text. Topic modelling might look like clustering of the text but it is different from it because in clustering we try to avoid the overlapping of the clusters but in topic modelling, one topic can be a part of multiple groups or a text group can have more than one topic as their feature. A representation of the most common words appearing in both the categories in concern is shown in Fig. 5. By topic modelling we try to find out the probable keywords or in the case of the Twitter platform the hashtags (#) which are mostly used in the case of irrelevant tweets so that a probable notification or alert about their inconsistent nature can be generated to the user before reading such tweets. This scheme in a way can help the user to avoid his dependency on tweets generating misinformation. For

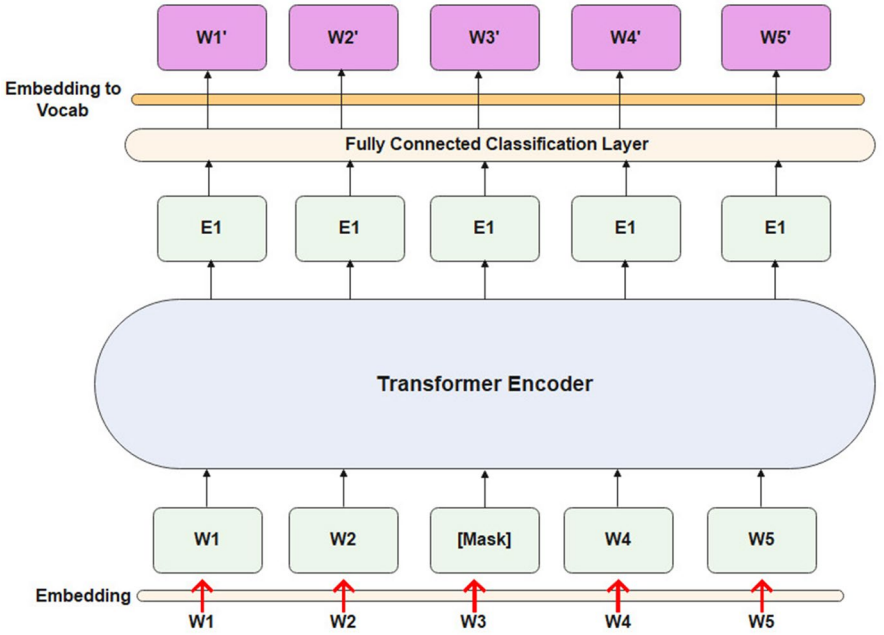

Fig. 4 BERT classification model

Fig. 5 Word cloud representing the frequent word in both relevant (left) and irrelevant (right) tweets

topic modelling, we used the two popular and widely used approaches which are latent Dirichlet allocation (LDA) [35] and latent semantic analysis (LSA) [36].

Latent Dirichlet allocation is a model based on probabilistic decisions which are used to carry out different operations such as judgment of similarity, classification, topic extraction over some discrete text corpus. The key feature of LDA is to view the document as a collection of different topics which analogues to the way humans perceive a document collection without any prior information. As we tend to categorize the things which seems to be identical in the same group and as we dig deep these groups can shuffle also, this is the same concept on which LDA extracts the topics from the text. The topic modelling is done based on two probability calculations, first the probability of assignment of a topic t for different words of document d, and next is the probability of a word w appearing in a topic t over the entire

document collection. Initially, k topics are assigned randomly to the different words and then an iterative process reassigns the words to different topics based on the above-discussed probabilities. For identification of the appropriate number of topics for our LDA model, we used the perplexity measure on a held-out test set [37]. The measure of perplexity decreases with the increasing likelihood of the LDA model so a model with a lower perplexity value is preferred. For our model, this perplexity curve goes flat after the topic value eight. Therefore, we selected the number of topics(k) = 8 as an optimal parameter for our dataset modelling.

Latent semantic analysis (LSA) or latent semantic indexing is a technique to determine the similarity of the text units based on the statistical measure of the likeliness of their meanings. The base of LSA is singular vector decomposition (SVD) using which it creates a vector space representation of the likelihood matrix. LSA is an application of SVD for the identification of semantic similarity between the text units based on the method of dimensionality reduction. SVD is a technique for dimensionality reduction of data applied to the matrix similar to principal component analysis (PCA) which is also a dimensionality reduction process based on eigenvalue decomposition. But the limitation of PCA was that it can be applied to only square matrices and SVD overcomes this limitation of PCA. The main theorem of SVD states that any matrix (A) can be decomposed into the product of three characteristic components namely left singular component (U), singular values (Σ) and right singular components (V^{T}) (Eq. 1):

$$A = U \cdot \Sigma \cdot V^{\mathrm{T}}. \tag{1}$$

The original matrix can be retrieved back by multiplying the components in order but for the sake of dimensionality reduction, the number of columns of the matrix U and the rows of matrix V can be controlled to reduce the dimension. This is called the k-rank approximation of SVD, where the U matrix represents the document-topic mapping and on the other hand the matrix V represents the term-topic mapping.

Result and Discussion

We first performed the classification of relevant tweets from non-relevant tweets using the standard machine learning approaches [38] and the word embedding models we used for this were Tf-IDF with unigram, bi-gram and tri-gram pairing and also word2vec. These are the machine learning models we used to perform the classification SVM, Random Forest, Decision tree, KNN, Naïve Bayes, LDA and Logistic Regression. Along with these we also used two ensemble learning-based algorithms namely ADA Boost and XG Boost. We compared the performance of these machine learning algorithms based on accuracy calculated by the confusion matrix the equation for which is shown as Eq. (2). The result of this comparison is shown in Table 4 along with accuracy we also calculated the area under the curve parameter of the ROC curve (receiver operating characteristic). The results have shown that the performance of the XGBoost algorithm was good with the highest accuracy of 79% (Indicated as bold in Table 4):

Table 4 Classification result comparison of machine learning algorithms with different word embedding models

Word-embedding	Algorithm	Accuracy	AUC	Time
Tf-idf1	SVM	0.78	0.59	5.74 s
Tf-idf1	Random Forest	0.77	0.55	34.2 s
Tf-idf1	Decision Tree	0.71	0.59	729 ms
Tf-idf1	KNN	0.75	0.59	496 ms
Tf-idf1	Naïve Bayes	0.61	0.57	1.29 s
Tf-idf1	LDA	0.58	0.57	35.2 s
Tf-idf1	Logistic Regression	0.76	0.55	547 ms
Tf-idf1	ADA Boost	0.75	0.52	2 min 29 s
Tf-idf1	**XG Boost**	**0.79**	0.59	1.97 s
Tf-idf2	SVM	0.77	0.57	14.6 s
Tf-idf2	Random Forest	0.76	0.54	1 min 43 s
Tf-idf2	Decision Tree	0.72	0.55	1.53 s
Tf-idf2	KNN	0.31	0.53	438 ms
Tf-idf2	Naïve Bayes	0.66	0.61	2.22 s
Tf-idf2	LDA	0.76	0.54	2 min 41 s
Tf-idf2	Logistic Regression	0.76	0.57	571 ms
Tf-idf2	ADA Boost	0.75	0.55	7 min 29 s
Tf-idf2	XG Boost	0.74	0.56	3.9 s
Tf-idf3	SVM	0.77	0.58	25 s
Tf-idf3	Random Forest	0.77	0.56	2 min 14 s
Tf-idf3	Decision Tree	0.67	0.54	2.92 s
Tf-idf3	KNN	0.27	0.5	777 ms
Tf-idf3	Naïve Bayes	0.73	0.63	7.98 s
Tf-idf3	LDA	0.75	0.56	4min16s
Tf-idf3	Logistic Regression	0.75	0.58	928 ms
Tf-idf3	ADA Boost	0.76	0.55	5 min 4 s
Tf-idf3	XG Boost	0.75	0.56	9.2 s
Word2Vec	SVM	0.66	0.63	20.7 s
Word2Vec	Random Forest	0.76	0.64	6.34 s
Word2Vec	Decision Tree	0.65	0.6	2.05 s
Word2Vec	KNN	0.43	0.6	2.76 s
Word2Vec	Naïve Bayes	0.55	0.59	336 ms
Word2Vec	LDA	0.67	0.64	718 ms
Word2Vec	Logistic Regression	0.67	0.63	617 ms
Word2Vec	ADA Boost	0.74	0.55	9 min 21 s
Word2Vec	XG Boost	0.75	0.56	5.2 s

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}). \quad (2)$$

Next, we performed the classification using a convolution neural network (CNN) [28], the parameters used and specifications of the CNN model are shown in Table 5.

Table 5 Tuning parameters for CNN classification model

Layer (type)	Output shape	Parameters
embedding (Embedding)	(None, 35, 100)	904,800
conv1d (Conv1D)	(None, 34, 32)	60
max_pooling1d (MaxPooling1D)	(None, 17, 32)	432
dropout (Dropout)	(None, 17, 32)	0
dense (Dense)	(None, 17, 32)	1056
dropout_1 (Dropout)	(None, 17, 32)	0
dense_1(Dense)	(None, 17, 16)	528
global_max_pooling1d (Global)	(None, 16)	0
dense_2 (Dense)	(None, 1)	17
Total params: 912,833 Trainable params: 912,833 Non-trainable params: 0		

Adam optimization algorithm is used and the loss function used in the model was binary cross-entropy as the classification is binary. The model was trained on 40 epochs and it achieved a validation accuracy of 83.1% after successful completion of the training.

Another method we used for the classification was based on the BERT model with a batch size of 64 and number of epochs = 10. One of the important parameters used to tune the BERT model is the learning rate, which has to be selected such as the model training loss should be minimum. The learning rate calculation is shown in Fig. 6. As it is evident from the figure that the loss is low when the learning rate is below 10^{-2}, and it is minimum at the value of 10^{-4}, so the same value will be used for training of our model. We already implemented SMOTE balancing to balance our dataset but to further diminish the effect of unbalancing we applied the *K*-fold cross-validation approach with stratified 5-folds. An equal number of samples for every fold was ensured by this stratification. The motive behind applying this strategy is to divide the dataset into 5 equal partitions and to use 4 of them as the training dataset and 1 as the validation dataset.

The training accuracy and loss along with the cross-validation accuracy and loss are shown in Fig. 7. The model performed well in both the training and testing phase

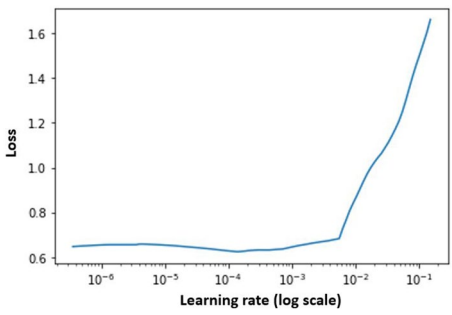

Fig. 6 Learning rate calculation of the BERT model

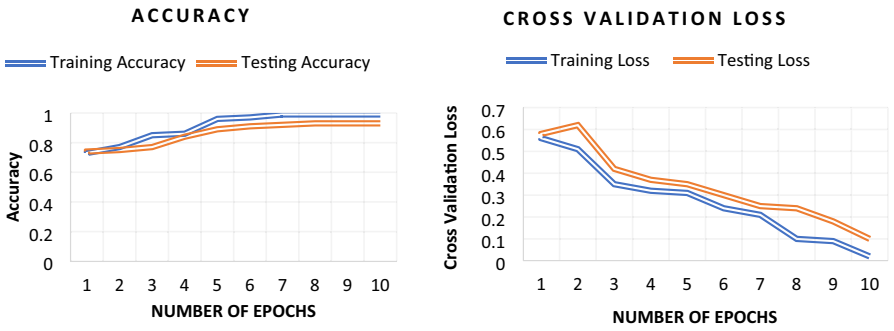

Fig. 7 Classification accuracy (left) and cross-entropy loss (right) using BERT

as visible from the figure, it achieved a validation accuracy of 92.8% at epoch level 10. The accuracy value becomes constant after epoch level 9 so a maximum of 10 epochs were used for calculation. All of the results were calculated on the Google Colab platform using the python programming language.

We compared the results of our different models with some of the state-of-the-art (SOTA) techniques present in the literature. The main objective of this comparison to the SOTA techniques is to make a reliable comparison although due to the unavailability of the datasets used in different SOTA techniques and novel characteristics of relevance in our dataset a straightforward comparison is not possible to the techniques available in the literature. Still, we carried out the comparison to those techniques which can be applied to our dataset without any modification and the comparison results based on classification accuracy, precision and recall are presented in Table 6. Out of the presented techniques, the proposed BERT model outperforms the other models by a huge margin (accuracy of 92.8%).

After classification we performed the topic modelling using the LSA and LDA approach, Fig. 8 represents the eight different topics found out and the three topmost frequent words used in that topic for the relevant tweets using LSA. Similarly, the topic modelling outcome using the LDA approach is shown in Fig. 9, with the most frequent words of each topic and the frequency of that topic. The topics modelled by LSA and LDA approach for the irrelevant tweets leading to misinformation are shown in Figs. 10 and 11, respectively. By looking at the topics formed by both the

Table 6 Comparison with the SOTA techniques

Model	Accuracy (%)	Precision	Recall
XGBoost	79	79.4	82.7
CNN [28]	83.1	84.2	78.3
Bonet-Jover et al. [6]	75	80.2	76.8
Pinto et al. [26]	79	80	84
Chakraborty et al. [38]	79	82.13	84.56
BERT (proposed)	92.8	93.6	91.2

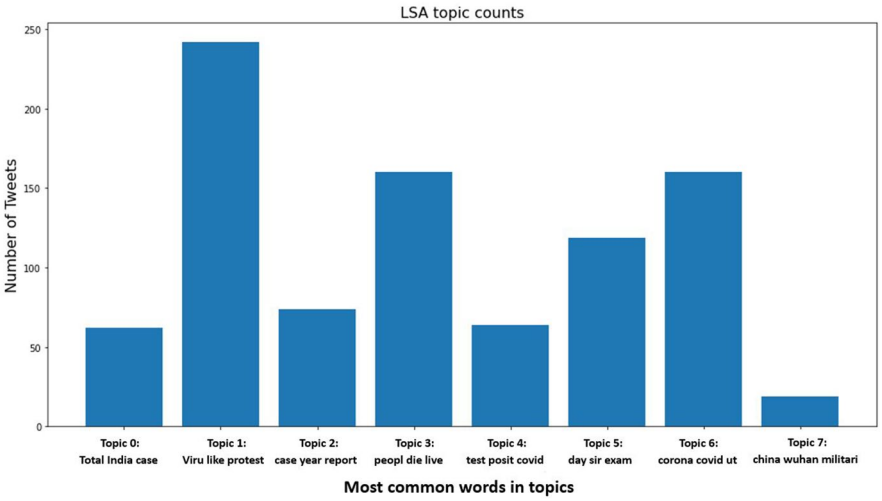

Fig. 8 Most common words in topics for relevant tweets using LSA

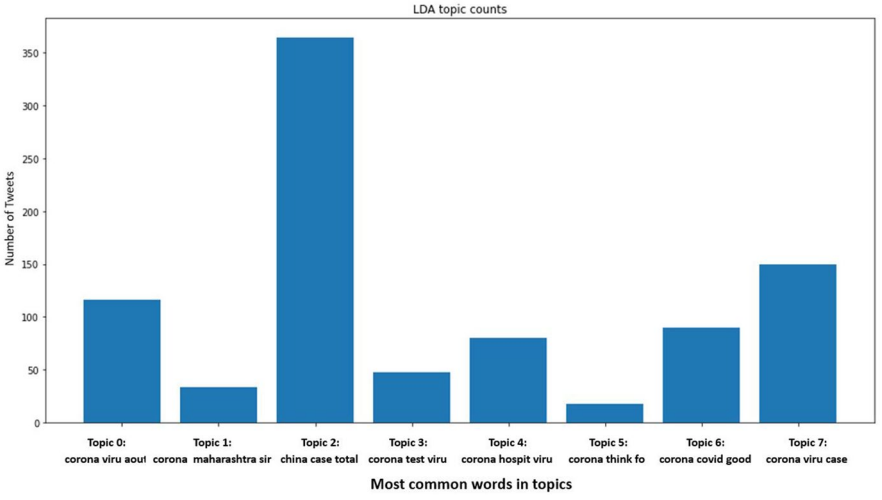

Fig. 9 Most common words in topics for relevant tweets using LDA

approaches, we can see that LDA approach models the topics more efficiently and in more precise manner.

To compare the performance of both the approaches we used the T-Distributed Stochastic Neighbour Embedding (t-SNE) clustering approach to visualize the topic clusters and to analyse their distribution. The outcome of the clustering approach is shown in Fig. 12 for the LSA approach and in Fig. 13 for the LDA approach. The t-SNE plot clearly distinguishes the better topic modelling approach in terms of clusters, it can be seen from the graphs that the clusters of topics formulated by the

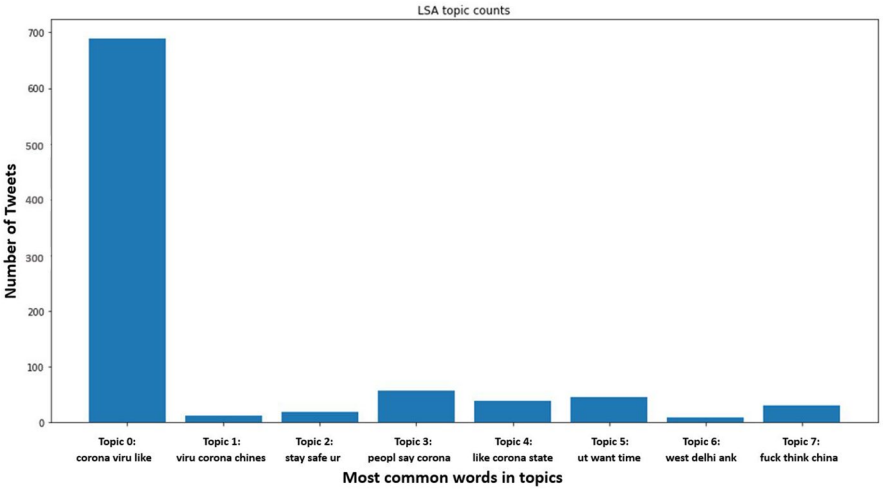

Fig. 10 Most common words in topics for irrelevant tweets using LSA

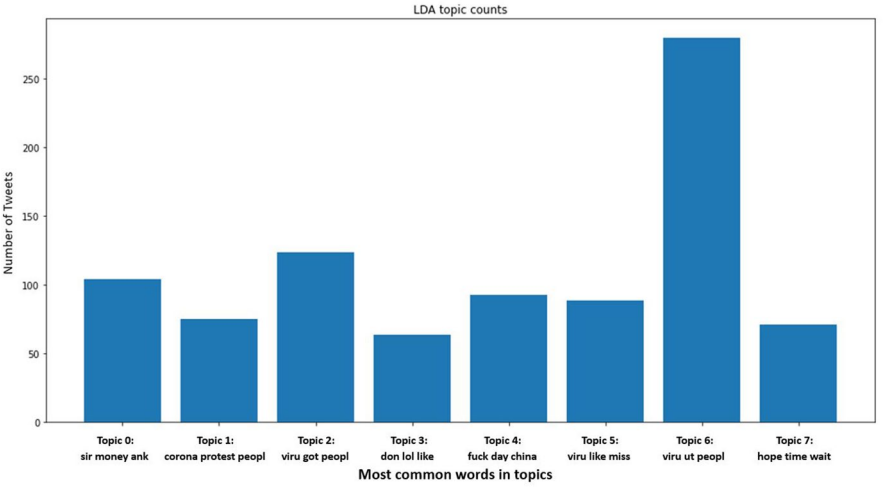

Fig. 11 Most common words in topics for irrelevant tweets using LDA

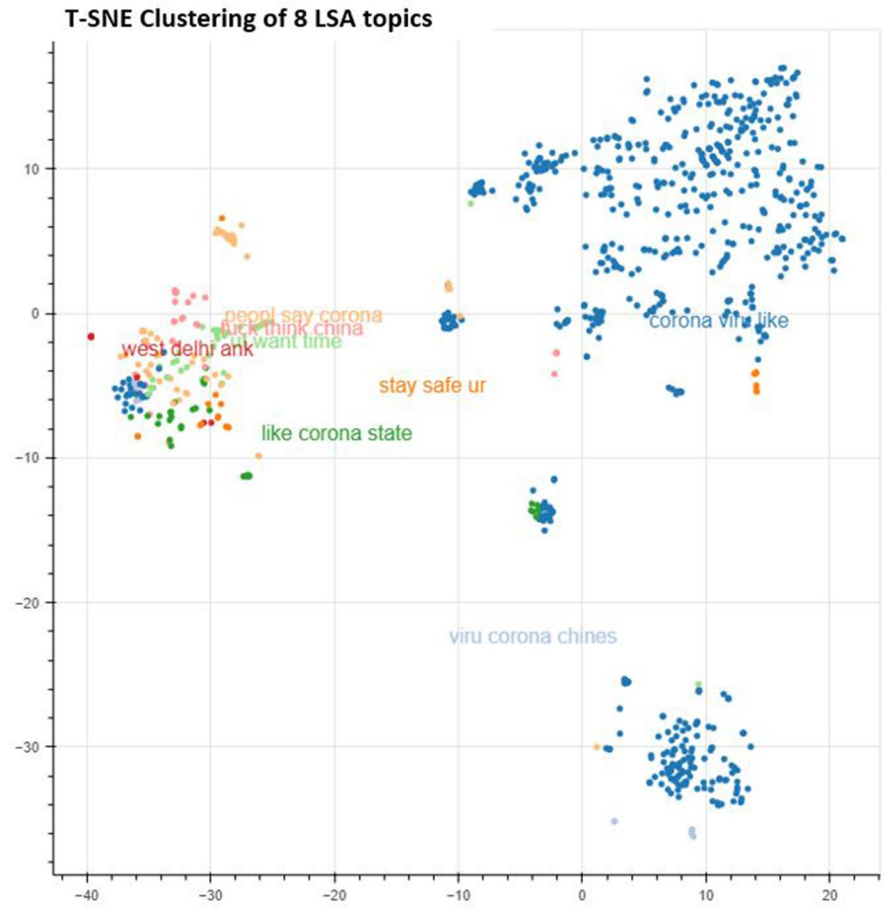

Fig. 12 t-SNE clustering of 8 topics using LSA approach

LDA approach are well defined and separated from each other while the cluster of topics formed by the LSA approach are overlapping and scattered in nature.

Conclusion and Future Scope

The paper elucidated the use of a transformer-based model in the classification of tweets on the basis of their relevance factor for the Covid-19 problem. The transformer-based model described in the paper was Bi-directional Encoder Representations from Transformers (BERT). Along with the presented model, the dataset was classified by machine learning algorithms with Tf-IDF and Word2Vec based vectorization approaches. The presented BERT-based model outperformed the existing state-of-the-art models based on accuracy with an accuracy of 92.8% for the

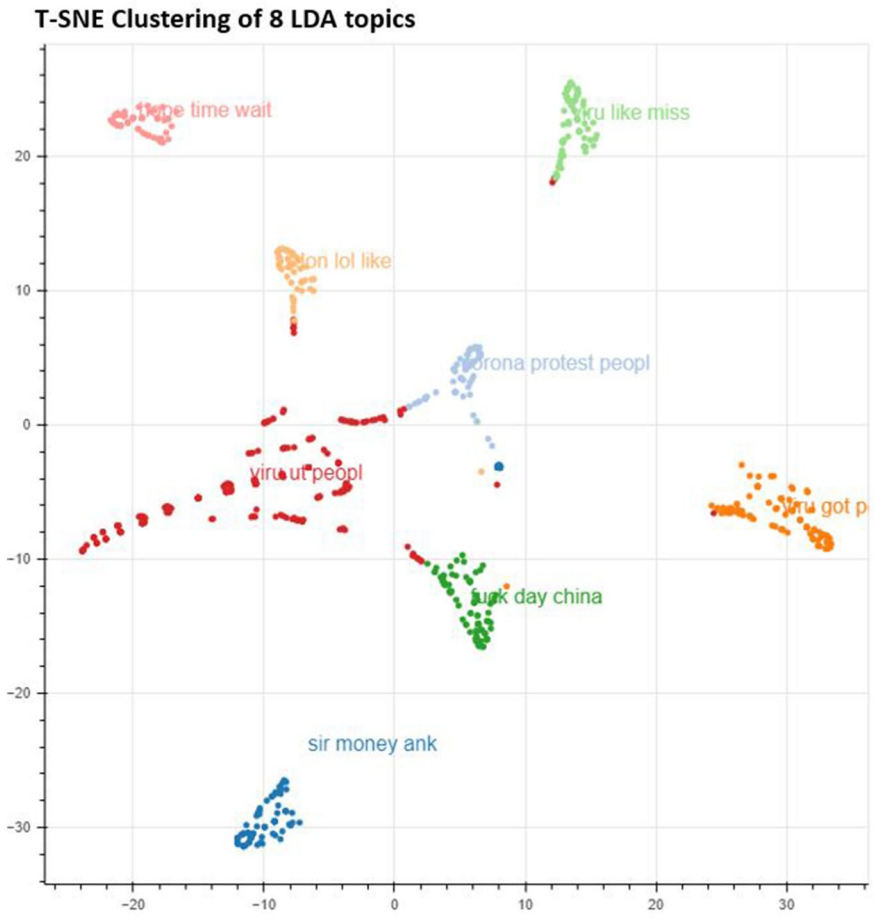

Fig. 13 t-SNE clustering of 8 topics using LDA approach

collected dataset. The work presented in the paper tries to solve the problem of classifying online social media contents based on their relativity to the topic for which the content is intended.

Along with the classification of relevant tweets the paper discussed the grouping of the keywords found in the relevant tweets in the form of topics. The words in the topics help the user to identify in the future that if a tweet contains a particular set of words, then that tweet can be considered relevant or not. For this topic modelling task, two approaches were presented in the paper namely LSA and LDA. The grouping of the keywords by the LDA approach was better as compared to the LSA approach.

This work constitutes the first step for the overall identification of facts vs fake news. After the classification of relevant tweets, the next step is to use only the relevant tweets to further classify them as fake or true. The irrelevant tweets are not

useful for the task of identification of facts as they mislead the topic of discussion. The dataset will also be enhanced by collecting more tweets over a different period of the pandemic to get a diversified view of the population.

References

1. Press Trust of India. https://www.hindustantimes.com/india-news/dhule-lynching-aftermath-rainpada-village-becomes-a-ghost-place/story-m3XjQYgS7uT0aDUc6ZJknM.html
2. Edosomwan, S., Prakasan, S.K., Kouamé, D., Watson, J., Seymour, T.: The history of social media and its impact on business. J. Appl. Manag. Entrep. **16**, 79 (2011)
3. Jin, S., Lin, W., Yin, H., et al.: Community structure mining in big data social media networks with MapReduce. Cluster Comput. **18**, 999–1010 (2015). https://doi.org/10.1007/s10586-015-0452-x
4. Schapals, A.K.: Fake news. J. Pract. **12**(8), 976–985 (2018). https://doi.org/10.1080/17512786.2018.1511822
5. University of Manchester.: New evidence shows might of Pharaoh Ramses is fake news. http://www.manchester.ac.uk/discover/news/new-evidence-shows-might-of-pharaoh-ramses-is-fake-news (2018)
6. Bonet-Jover, A., Piad-Morffis, A., Saquete, E., Martínez-Barco, P., García-Cumbreras, M.Á.: Exploiting discourse structure of traditional digital media to enhance automatic fake news detection. Expert. Syst. Appl. (2020). https://doi.org/10.1016/j.eswa.2020.114340. (**ISSN 0957-4174**)
7. Fan, C., Wu, F., Mostafavi, A.: A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. IEEE Access **8**, 10478–10490 (2020). https://doi.org/10.1109/ACCESS.2020.2965550
8. Cepoi, C.-O.: Asymmetric dependence between stock market returns and news during COVID-19 financial turmoil. Financ. Res. Lett. (2020). https://doi.org/10.1016/j.frl.2020.101658. (**ISSN 1544-6123**)
9. Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L.: The science of fake news. Science **359**(6380), 1094–1096 (2018). https://doi.org/10.1126/science.aao2998
10. Talwar, S., Dhir, A., Singh, D., Virk, G.S., Salo, J.: Sharing of fake news on social media: application of the honeycomb framework and the third-person effect hypothesis. J. Retail. Consum. Serv. (2020). https://doi.org/10.1016/j.jretconser.2020.102197. (**ISSN 0969-6989**)
11. Obiała, J., Obiała, K., Mańczak, M., Owoc, J., Olszewski, R.: COVID-19 misinformation: accuracy of articles about coronavirus prevention mostly shared on social media. Health Policy Technol. (2020). https://doi.org/10.1016/j.hlpt.2020.10.007. (**ISSN 2211-8837**)
12. Cherry, J.D.: The chronology of the 2002–2003 SARS mini pandemic. Paediatr. Respir. Rev. **5**(4), 262–269 (2004). https://doi.org/10.1016/j.prrv.2004.07.009. (**ISSN 1526-0542**)
13. Lin, C.-Y., Broström, A., Griffiths, M.D., Pakpour, A.H.: Investigating mediated effects of fear of COVID-19 and COVID-19 misunderstanding in the association between problematic social media use, psychological distress, and insomnia. Int. Intervent. (2020). https://doi.org/10.1016/j.invent.2020.100345. (**ISSN 2214-7829**)
14. Ceron, W., de-Lima-Santos, M.-F., Quiles, M.G., : Fake news agenda in the era of COVID-19: identifying trends through fact-checking content. Online Soc. Netw. Med. **21**, 100116 (2021). https://doi.org/10.1016/j.osnem.2020.100116
15. Zika virus. https://www.who.int/news-room/fact-sheets/detail/zika-virus
16. Ebola virus disease. https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease
17. Middle east respiratory syndrome corona virus. https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers#tab=tab_1
18. Shin, S.Y., Seo, D.W., An, J., et al.: High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep **6**, 32920 (2016). https://doi.org/10.1038/srep32920
19. Sommariva, S., Vamos, C., Mantzarlis, A., Đào, L.-L., Tyson, D.M.: Spreading the (fake) news: exploring health messages on social media and the implications for health professionals using a case study. Am. J. Health Educ. **49**(4), 246–255 (2018). https://doi.org/10.1080/19325037.2018.1473178

20. Sun, K., Chen, J., Viboud, C.: Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: a population-level observational study. Lancet Digit. Health **2**(4), e201–e208 (2020). https://doi.org/10.1016/S2589-7500(20)30026-1. (**ISSN 2589-7500**)
21. Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C.S., Ho, R.C.: Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. Int. J. Environ. Res. Public Health **17**, 1729 (2020)
22. Guntuku, S.C., Sherman, G., Stokes, D.C., et al.: Tracking mental health and symptom mentions on twitter during COVID-19. J. Gen. Intern. Med. **35**, 2798–2800 (2020). https://doi.org/10.1007/s11606-020-05988-8
23. Bastick, Z.: Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. Comput. Hum. Behav. (2021). https://doi.org/10.1016/j.chb.2020.106633. (**ISSN 0747-5632**)
24. Kwon, J., Grady, C., Feliciano, J.T., Fodeh, S.J.: Defining facets of social distancing during the COVID-19 pandemic: Twitter analysis. J. Biomed. Inform. (2020). https://doi.org/10.1016/j.jbi.2020.103601. (**ISSN 1532-0464**)
25. Singh, M., Jakhar, A.K., Pandey, S.: Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc. Netw. Anal. Min. **11**, 33 (2021). https://doi.org/10.1007/s13278-021-00737-z
26. Pinto, A., Gonçalo Oliveira, H., Figueira, Á., et al.: Predicting the relevance of social media posts based on linguistic features and journalistic criteria. New Gener. Comput. **35**, 451–472 (2017). https://doi.org/10.1007/s00354-017-0015-1
27. Rudra, K., Sharma, A., Ganguly, N., et al.: Classifying and summarizing information from microblogs during epidemics. Inf. Syst. Front. **20**, 933–948 (2018). https://doi.org/10.1007/s10796-018-9844-9
28. Madichetty, S., Sridevi, M.: Detecting informative Tweets during disaster using deep neural networks. In: 2019 11th International Conference on Communication Systems and Networks (COMSNETS), 2019, pp. 709–713. https://doi.org/10.1109/COMSNETS.2019.8711095
29. Bhoi, A., Pujari, S.P., Balabantaray, R.C.: A deep learning-based social media text analysis framework for disaster resource management. Soc. Netw. Anal. Min. **10**, 78 (2020). https://doi.org/10.1007/s13278-020-00692-1
30. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
31. Eklund, M.: comparing feature extraction methods and effects of pre-processing methods for multilabel classification of textual data (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-231438 (2018)
32. Liu, Y., Ju, S., Wang, J., Su, C.: A new feature selection method for text classification based on independent feature space search. Math. Probl. Eng. **2020**, 6076272 (2020). https://doi.org/10.1155/2020/6076272
33. Tomas, M., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. http://arxiv.org/abs/1301.3781 (2013)
34. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805 [cs.CL] (2018)
35. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(2), 993–1022 (2003)
36. Berry, M., Young, P.: Using latent semantic indexing for multilanguage information retrieval. Comput. Humanit. **29**(6), 413–429 (1995)
37. Slof, D., Frasincar, F., Matsiiako, V.: A competing risks model based on latent Dirichlet allocation for predicting churn reasons. Decis. Support Syst. (2021). https://doi.org/10.1016/j.dss.2021.113541. (**ISSN 0167-9236**)
38. Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A.E.: Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—a study to show how popularity is affecting accuracy in social media. Appl. Soft. Comput. (2020). https://doi.org/10.1016/j.asoc.2020.106754. (**ISSN 1568-4946**)

Authors and Affiliations

Utkarsh Sharma[1] · Prateek Pandey[1] · Shishir Kumar[2]

[1] Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna, India

[2] Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India