Leveraging Transformer Based Approach to Improve Rumor Classification From Twitter Data

Abstract

For obtaining real-time information, general people and traditional news media's dependency on social media posts are constantly rising. This dependency has a direct relation to rumor propagation. If the veracity of a piece of information is known to be false, people show a low inclination towards sharing that information. In order to mitigate the rumor's negative effect on not only the individuals but also on the community, society, and country, a self-regulating, low-cost, and futuristic rumor detection model is required and a few types of research have already been conducted. With this goal in mind, we have presented a novel model which incorporates three sturdy machine learning-based models named XGBoost, Support Vector Machine (SVM), and Random Forest as well as two deep learning-based transformers called BERT and DistilBert. Our proposed model has been trained and tested on a merged dataset of twitter15 and twitter16 for the machine learning model, SVM has achieved the best accuracy of 87.56%. On the other hand, for the learning-based transformers, BERT has achieved the best accuracy of 90.20%. To conclude, these have effectively surpassed all its competitors in terms of accuracy, recall, precision, and f1-Score.

Introduction

Due to the rapid developments in the structure and accessibility of handheld devices and the internet, people's presence, usage, and dependency on Online Social Networks (OSNs) have significantly increased. People feel more connected to each other as social networking sites function as a communication medium between different groups, organizations, and individuals despite their geographical location in the world. The structural characteristics and the openness of the OSNs allow numerous users to effortlessly access and share any information on any social media (e.g., Facebook, Reddit, Youtube, Twitter, and Sina Weibo). This shared information can be in any cross-media form such as text, image, audio, and video. Due to constant real-time information shared by the users, general people, and even reporters, often seek fresh updates on an emergency event or breaking news from social media. The information posted in this situation is often not verified due to the lack of time which leads to the creation of rumors. Several definitions of rumor exist (Veyseh et al. 2019; Guo et al. 2018; Santhoshkumar and Dhinesh Babu 2020; Kumar and Sangwan 2019). Any information whose authenticity has not been confirmed at the time of posting is generally considered a rumor.

Rumors are created unintentionally as well as deliberately. Deliberately created rumors are often used to shape public opinions regarding a focused event, an individual, or an organization where the main motive is personal, professional, political, or economic gain (Hosni, Li, and Ahmad 2020; Ma et al. 2020). Additionally, a few deliberately created rumors are focused on entertainment and sarcasm where the main purpose is amusement. On the contrary, unintentional rumors are created and disseminated, in the cases of natural calamities (e.g. cyclones earthquakes, and floods) or emergency situations (e.g. terrorist attacks, building collapse, fire incidents, and plane crashes) when there is a scarcity of real-time information from the traditional news media (Ahsan, Kumari, and Sharma 2019; Pathak et al. 2020). Lack of information causes anxiety among the people. Anxiety reduces the cognitive ability of individuals and thus they are more inclined towards social networking sites to look and share relevant information associated with the event ranging over different social media platforms without checking the veracity of the information (Asghar et al. 2021). In addition, due to easy and low-cost access to the internet as well as hand-held devices, users without any professional journalism knowledge share their opinion or self-observed information directly on social sites without prior fact-checking which prevails in the emergence of rumors (Ma et al. 2020). Examples of a few recent rumors are:

- When the first batch of COVID-19 vaccinations was given in the U.S in 2020, a photoshopped image of a CNN report claiming the vaccinations turned the patients into cannibals was circulating among various social media sites. This image spread unwanted anxiety and fear among the patients who are or will be vaccinated (Mikkelson 2021).
- In 2018, a rumor circulated on the social media that the broiler chicken is a source of life-threatening Nipah virus which caused the heavy economical loss, especially to the people of Tamil Nadu (Pathak et al. 2020)
- The widespread rumors of problems associated with the vaccinations of measles, mumps, and pertussis originated the "Anti-vax" movement. This signifi-

cantly escalated the number of people affected with diseases easily preventable by vaccines, as well as their associated death rate (Benecke and DeYoung 2019).

- Throughout the 2016 American Presidential election, about 529 rumors were circulating on various OSNs which had a serious impact on the voting results (Veyseh et al. 2019).

Due to rumor's huge negative impact on individuals, groups, or society at large, several types of research have been undertaken by the academics and industry experts encompassing the fields of psychology, humanities, political science, management science, and computer science (Ahsan, Kumari, and Sharma 2019; Chen et al. 2019). In most OSNs, a sophisticated filtering mechanism for checking the veracity of the posted contents is overlooked (Asghar et al. 2021). Thus they are dependent on their user's reports for identifying rumors that further have to be checked and confirmed by professionals. This verification requires a notable amount of manual labor, economical support, and time. Within this verification period, owing to the layout of OSNs those rumors remain to be disseminated (Veyseh et al. 2019; Asghar et al. 2021; Chen et al. 2019). Additionally, for debunking rumors various fact-checking websites are available. For example snopes.com, emergent.info, twittertrails.com, factcheck.org, and politifact.com (Veyseh et al. 2019; Pathak et al. 2020; Chen et al. 2019). These websites often suffer from low coverage of topics and inefficiency (Chen et al. 2019; Ma, Gao, and Wong 2019). For solving the above-mentioned problems an automatic, economical, and advanced rumor detection model with a short response time is needed. Several types of research have been carried out for the detection of rumors (Ahsan, Kumari, and Sharma 2019; Pathak et al. 2020; Santhoshkumar and Dhinesh Babu 2020). For this purpose, we have formed a novel model. Three state-of-the-art, sturdy machine learning-based models (XGBoost, Support Vector Machine (SVM), and Random Forest) and two deep learning-based transformers (BERT and DistilBert) were incorporated into our model.

Related Work

The number of people on multiple online social networks (OSNs), as well as their involvement in the proliferation of rumors, has soared exponentially. Rumors are the contents of a post or replies to any particular post published and shared by one or multiple social media users without proper verification of the truthfulness of the content while posting. Due to the open, simple, and unrestricted nature of social media platforms and their widespread access, people often spread misinformation or rumors without taking into consideration the reliability of the data that might be detrimental to an individual, group, government, or even an entire society. For retaining the health of the online social media environment, a considerable amount of research has been conducted on the detection and mitigation of the propagation of rumors.

Several strategies have been undertaken for this purpose which includes the dissemination of correct information to counteract the effect of rumors (Yang, Li, and Giua 2020), identification of the rumor source (Shelke and Attar 2019), understanding the influence of a rumor on an individual (Hosni, Li, and Ahmad 2020), and his/her viewpoint towards the rumor (Hosni, Li, and Ahmad 2020), the semantic relationship and pattern between the original post and their replies in the context of rumor detection (Veyseh et al. 2019). Lastly, exploration of the individual behaviors as well as the social behaviors in OSNs has also aided in lessening the proliferation of rumors (Hosni, Li, and Ahmad 2020).

Researchers and industry experts took several approaches for detecting rumors, including supervised and unsupervised Learning, Deep Learning, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), and hybrid approaches (Kumar and Sangwan 2019). Kumar et al. in have surveyed all the machine learning-based techniques used for the classification of rumor and not-rumors and have identified the most popular ones which include SVM, Naive Bayes, and Decision Trees. However, owing to the high success rate of rumor detection, the researchers are now more inclined toward Deep Learning approaches. In (Santhoshkumar and Dhinesh Babu 2020), for the early detection of the rumors, a certainty factor-based, dual convolutional neural network (DCNN) leveraging the inherent features of information has been proposed which can successfully detect rumors with a very insignificant number of posts (sparse data). The authors (Chen et al. 2019) proposed an attention-residual network merged with CNN that utilized the latent contextual features as well as three fine-tuned attention mechanism models to seize the long-range relation of the dependency for rumor detection. In (Veyseh et al. 2019), a multi-task learning framework was introduced to prioritize the main post instead of the replies of the posts to exploit the semantic relations between the contents of the posts.

Asghar et al. (Asghar et al. 2021) proposed a deep learning-based model named BiLSTM-CNN that uses the bidirectional LSTM layer (BiLSTM) layer to assimilate the long-term relation of a post's past and future context information. Then a feature extraction is conducted using the CNN model which helps in the classification of rumor and not-rumor. The RNN-based rumor detection mechanism presented by (Ma et al. 2020), uses two variants of RvNN models to better integrate the information of structural and textual properties existing in the posts. Ma et al. designed a text-based framework using Generative Adversarial Networks (GAN) which utilizes a co-dependent text generator and discriminator for the recognition of rumor characteristics from a post (Ma, Gao, and Wong 2019).

Methodology

In figure 1, we can see the flow of our work. We started with collecting the data for our research. Afterward, we applied several data-preprocessing techniques to best fit our dataset into different machine learning models. However, in the very next phase, we had either split out the dataset into training and testing sets or first of all done tokenization and then split our dataset into training and testing sets. As a result, we have actually generated two different groups of training-testing pair sets to train the machine learning models like-XGBoost, SVM, Random Forest, and transformer models like- BERT, and DistilBERT. Furthermore, in the very next step, we fed the testing sets into the trained machine learning models and trained transfer models respectively. Last of all, we evaluated the model performance using different metrics and performed a result analysis to figure out the best-suited models.

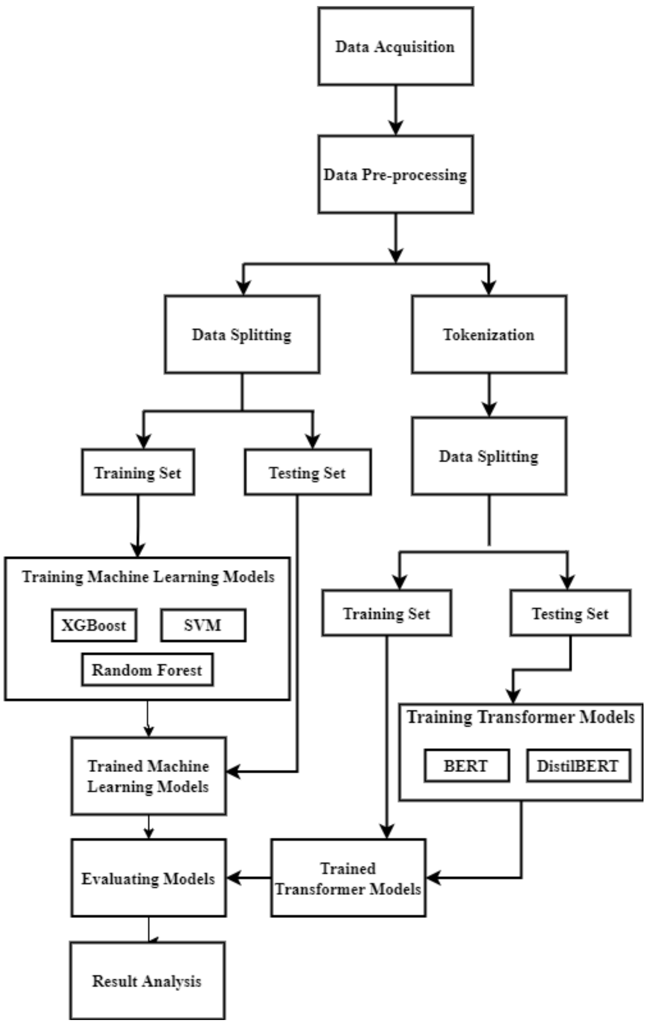

Figure 1: Workflow diagram

Data Acquisition

Data acquisition is one of the most crucial steps for rumor detection which is often challenging due to the lack of data extraction facilities. For example, Twitter allows data to be extracted that are a maximum of 7 days old. So, to access data older than that web crawlers might be used. On the other hand, maximum user profiles on Facebook only share data with their inner circle that has been associated as “friends”. So, the Facebook data mostly consists of data from public pages and public user profiles (Kumar and Sangwan 2019). For training our model and result generation, we have used open-source datasets Twitter15 (Liu et al. 2015) and Twitter16 (Ma et al. 2016).

Data Pre-processing

Preprocessing data is an important step in the building of a Machine Learning model, and the results are determined by how well the data is preprocessed. It lays the groundwork for valid data analysis by removing errors from the data. Because of the inherent complexity of building operations and the lack of data quality, it is an absolutely necessary phase in building operational data analysis. In data science, data preprocessing is a collection of strategies for improving the quality of raw data. (Fan et al. 2021)

Tokenization

The procedure of separating words and idioms from a text into individual words or pairs of words by following criteria set by the researchers is usually known as Tokenization and the separated words are acknowledged as Tokens. It aids us in the comprehension of the context as well as the creation of the NLP mode by providing crucial information encoded in the order of words. For example, nouns and pronouns are generally trailed by a verb. (Chakravarthy 2020)

Data Spliting

For the Machine Learning models, our merged dataset was split into a ratio of 7.5:2.5 where the train set comprised 75% of the data and the test set comprised 25% of the data. Though for the Deep Learning-based transformers, the combined dataset was split into a ratio of 8:2 where the train set comprised 80% of the data and the test set comprised 20% of the data.

Dataset Description

For mitigating the adverse effect and impact of rumor propagation in OSNs, we first need to effectively classify rumors and non-rumors and therefore we need confirmed labeled data ranging over a variety of subjects.

Figure 2: Twitter15 dataset topics

For this purpose, we utilized publicly available Twitter15 (Liu et al. 2015) and Twitter16 (Ma et al. 2016) datasets which have labeled samples that have been further verified using rumor-exposing websites such as snopes.com, emergent.info, twittertrails.com, factcheck.org, and politifact.com (Veyseh et al. 2019; Pathak et al. 2020; Chen et al. 2019).

Figure 3: Twitter16 dataset topics

Figure 2 and figure 3 illustrate the variety of subject matter of Twitter15 (Liu et al. 2015) dataset and Twitter16 (Ma et al. 2016) datasets respectively. In addition to this, the Twitter15 dataset was also validated manually by three researchers (Liu et al. 2015).

Table 1: Statistics of datasets

Label	Twitter15	Twitter16	Merged Dataset
Rumor	746	412	1158
Non-rumor	370	205	575
Total texts	**1116**	**617**	**1733**

Statistics of both the datasets and the merged dataset have been presented in table 1. The merged dataset had a total of 1733 samples among those 1158 were rumors and 575 were non-rumors. Similar to (Tafannum et al. 2021), we merged both datasets for being used in our model.

Table 2: Accumulated samples from Twitter15 & Twitter16 Dataset

Dataset	Label	Text
Twitter15	FALSE = rumor	42 million dead in bloodiest Black Friday weekend on record
	TRUE = non-rumors	uniformed Canadian soldier shot at the war memorial in #ottawa.
Twitter16	FALSE = rumor	In heaven chick-fil-a is open on Sundays.
	TRUE = non-rumors	white house lit in rainbow colors after gay marriage ruling

Table 2, shows the samples of rumored and non-rumored posts of both Twitter15 and Twitter16 datasets.

Our Used Models

In our study, we have used three machine learning-based models such as XGBoost, Support Vector Machine (SVM), and Random Forest along with two deep learning-based transformers such as BERT and DistilBert.

XGBoost: XGBoost is a well-known and convenient open-source version of the gradient boosted trees algorithm, and it's free. The algorithm Gradient Boosting is used to try to figure out what a target variable is going to be by merging the estimates of a group of simpler, weaker models to get a more accurate prediction (Chen and Guestrin 2016).

Support Vector Machine (SVM): It performs by mapping data to a high-dimensional classifier, which makes it possible to categorize the data points, even if the data aren't linearly separable. This is how SVM works. People look for a way to separate the different categories. Then, the data are changed so the hyperplane that separates them can be drawn. New data can then be used to figure out which group a new record should belong to (Evgeniou and Pontil 1999).

Random Forest: Employing ensembles of trees, where each tree in the ensemble is developed according to a random parameter, can yield significant gains in classification and regression accuracy. The ensemble's final predictions are obtained by aggregating them. These approaches are known as "random forests" because the ensemble's basic constituents are tree-structured predictors, and each of these trees is generated using an injection of randomness (Breiman 2001).

BERT: BERT is an acronym for Bidirectional Encoder Representations. It is one of the open-sourced machine learning frameworks currently available that is frequently used in NLP. Its main purpose is the comprehension and disambiguation of ambiguous words present in a text by analyzing the context of the surrounding text. Additionally, deep bidirectional representations are pre-trained by conditioning on both the left and right context simultaneously (Devlin et al. 2018).

DistilBERT: DistilBERT is a Transformer model that is compact, fast, inexpensive, and light, and it is developed by distilling BERT basis. In comparison to bert-base-uncased, it has 40 percent less parameters and runs 60 percent faster while retaining more than 95 percent of BERT's performance on the GLUE language understanding benchmark (Sanh et al. 2019).

Result Analysis

For the performance evaluation of our model, we used several criteria such as accuracy score, precision score, recall score, confusion matrix, ROC curve, and Fbeta measure.

Accuracy: The ratio of correct predictions to total sampled data is known as accuracy.

$$Accuracy = \frac{Number\ Of\ Correct\ Prediction}{Total\ Number\ Of\ Predictions\ Made} \tag{1}$$

Precision: Precision refers to the accuracy of the classifiers. It belongs to the [0, 1] range. It is computed as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2}$$

Recall: Recall is a metric that indicates how comprehensive a classifier is. It belongs to the [0, 1] range, and is computed as follows:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3}$$

f1-Score: The Harmonic Mean of accuracy and recall is represented by the f1-Score, which has a range of [0, 1]. It indicates the accuracy of the classifier. A higher f1-Score indicates that our model is more accurate. It may be expressed numerically as follows:

$$f1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Confusion Matrix: The confusion matrix is a prominent metric for addressing classification difficulties. It may be used for both binary classification and multiclass classification issues. Table 3 shows an example of a confusion matrix for binary classification.

Table 3: An Example of Confusion Matrix

Labels	Negative	Positive
Negative	True Negative (TN)	False Positive (FP)
Positive	False Negative (FN)	True Positive (TP)

The sum of predicted and actual counts is represented by confusion matrices. The number of correctly detected negative situations is represented by the output True Negative. Similarly, True Positive represents the number of positively detected occurrences that were accurately identified. The False Positive is the number of real negative cases labeled as positive, whereas the False Negative is the number of genuine positive cases categorized as negative.

ROC Curve A receiver The operational characteristic curve (ROC curve) is a graph that illustrates a classification model's performance across all categorization levels. This graph shows two parameters named True Positive Rate and False Positive Rate.

The ROC curve that is closest to the top left corner of the figure indicates how well the model categorizes the data. We calculated the AUC (area under the curve) to see how much of the plot sits below the curve.

Fbeta-measure:

The Fbeta-measure is an abstraction of the F-measure, in which a coefficient called beta controls the balance between accuracy and recall in the computation of the harmonic mean.

Result Analysis for Machine Learning Models

Here in this section, we have analysed the results from the machine learning models in terms of their accuracy, precision, racall, f1-score by varying the beta value, ROC curve and the confusion matrix.

Classification Report

The figure 4 shows the results after applying the classification algorithms to the test dataset.

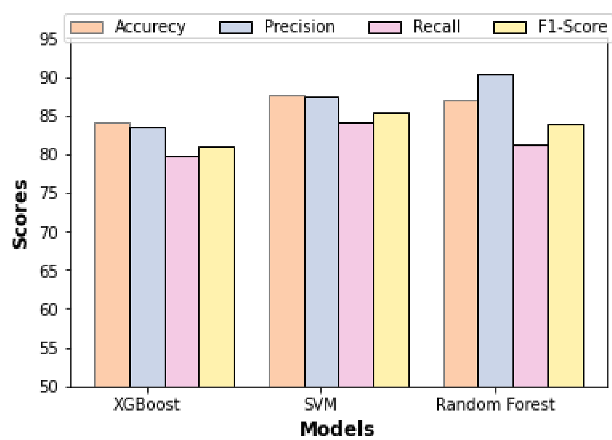

Figure 4: Classification Report of Machine Learning Models

From the accuracy provided in figure 4, we can see that XGBoost, Support Vector Machine (SVM), and Random Forest Classifier achieved 84.10%, 87.56% and 87.10% of accuracy. So, we can determine that With an accuracy rate of 87.56% and a weighted average f1-score of 85.36%, Support Vector Machine produces the greatest results. We created confusion matrices, Fbeta measure, and ROC curves for the algorithms to further test their performance and determine the findings.

Confusion matrix

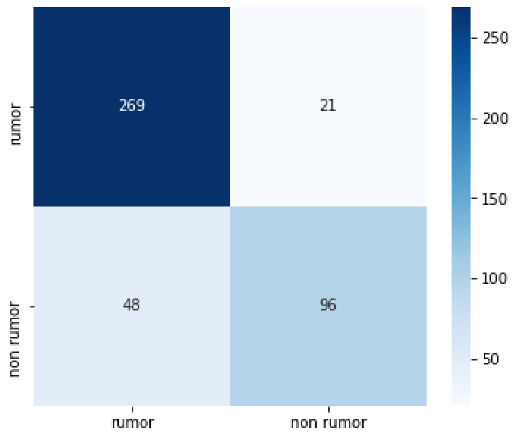

Figure 5: Confusion matrix of XGBoost

From figure 5 we can see that the confusion matrix of XG-Boost predicted (269+96) or 365 test instances accurately and (21+48) or 69 test instances incorrectly.

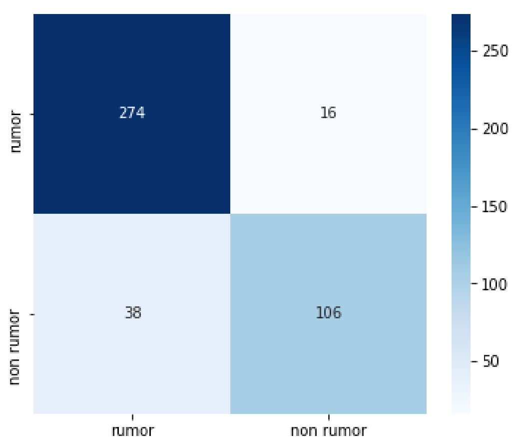

Figure 6: Confusion matrix of SVM

From figure 6 we can see that the confusion matrix of Support Vector Machine (SVM) predicted (274+106) or 380 test instances accurately and (16+38) or 54 test instances incorrectly.

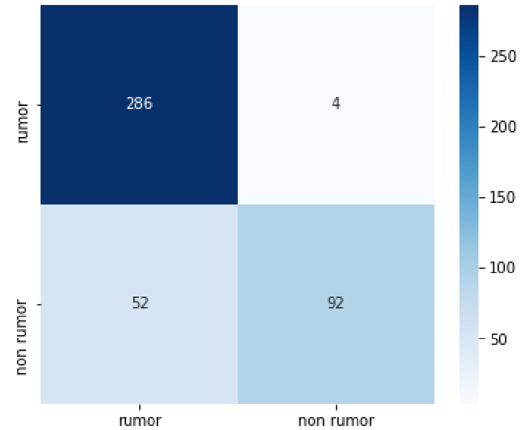

Figure 7: Confusion matrix of Random Forest

From figure 7 we can see that the confusion matrix of Random Forest predicted (286+92) or 378 test instances accurately and (4+52) or 56 test instances incorrectly.

Fbeta Measure

In this section, we have analysed the f1-score for different beta (0.5,1.0,2.0) values.

Table 4: Fbeta measure of machine learning models

Models	Fbeta(=0.5)	Fbeta(=1)	Fbeta(=2)
XGBoost	0.7843	0.7356	0.6926
SVM	0.8386	0.7970	0.7593
Random Forest	0.8712	0.7667	0.6845

From Table 5, we can see that for the three models the false positive predictions are high in number. As maximizing precision minimizes false positives, we need to consider the Fbeta(=0.5)-measure from table 4 that puts more attention on minimizing false positives than minimizing false negatives.

Table 5: False predictions of machine learning models

Models	False Positive	False Negative
XGBoost	48	21
SVM	38	16
Random Forest	52	4

ROC curve

From figure 8 we found that SVM has an AUC value of 0.84 that indicates that this is the best out of the three machine learning models.

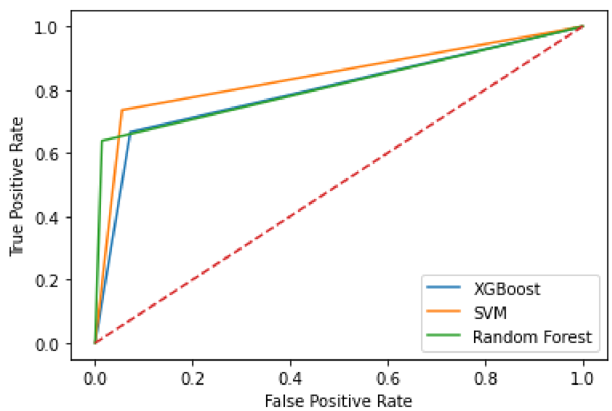

Figure 8: ROC Curve of machine learning models

Results from the Deep Learning Models

Here in this section, we have analysed the results from the deep learning models in terms of their training curves, testing curves, accuracy, precision, racall, f1-score by varying the beta value, ROC curve and the confusion matrix.

Training and Testing Curves

The assessment metrics for assessing the performance of BERT and DistilBERT are training and validation accuracy and training and validation loss.

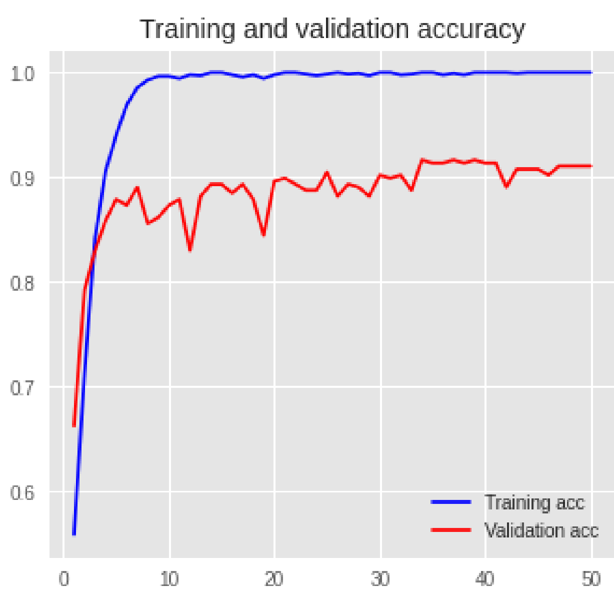

Figure 9: Training and Validation Accuracy Curves of BERT Model

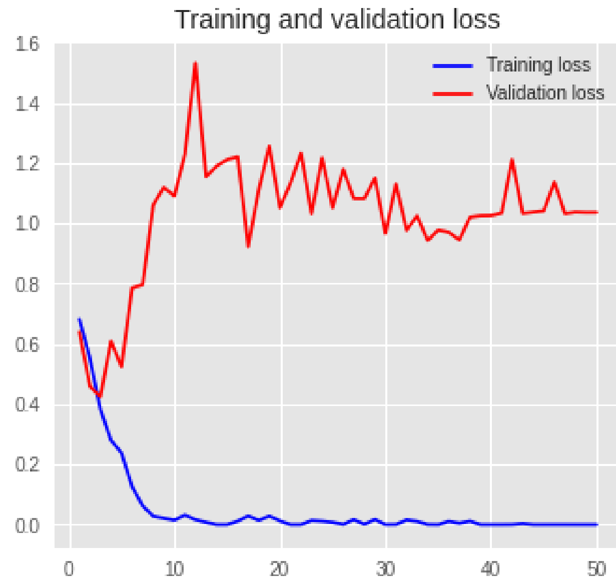

Figure 10: Training and Validation Loss Curve of BERT Model

In figure 9 and figure 10, Validation loss reduces as validation accuracy increases in our BERT model, indicating that the model is learning properly and achieving a 90.20% overall accuracy.

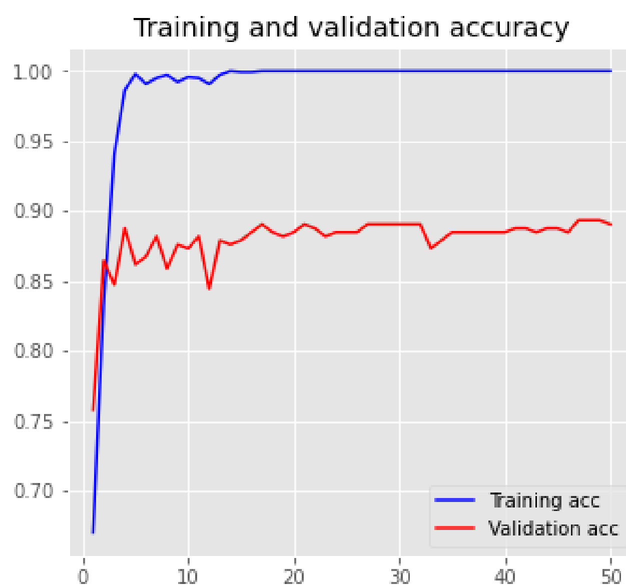

Figure 11: Training and Validation Accuracy Curves of DistilBERT Model

On the otherhand, figure 11 and figure 12 show that Validation loss reduces as validation accuracy increases in our DistilBERT model, indicating that the model is learning properly and achieving a 89.05% overall accuracy

.

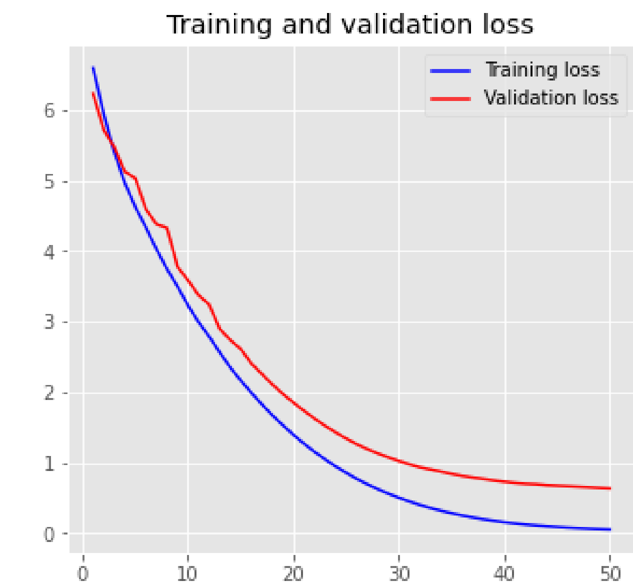

Figure 12: Training and Validation Loss Curve of DistilBERT Model

Classification Report

The figure 4 shows the results after applying the classification algorithms to the test dataset.

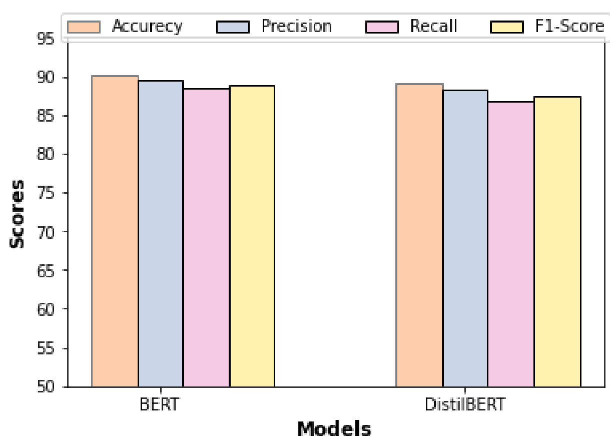

Figure 13: Classification Report of Machine Learning Models

From the accuracy provided in figure 13, we can see that BERT model and DistilBERT model have achieved 90.20% and 89.05% of accuracy respectively. Here, the BERT model shows us 89.46% of precision score, 88.41% of recall score and 88.90% of f1-score. On the other hand, the DistilBERT model shows us 88.24% of precision score, 86.84% of recall score and 87.48% of f1-score. With an accuracy rate of 90.20% and a weighted average f1-score of 88.90%, BERT model produces the greatest results. We created confusion matrices, ROC Curve and Fbeta measure for the algorithms to further test their performance and determine the findings.

Confusion matrix

From figure 14 we can see that the confusion matrix of BERT model predicted (216+97) or 313 test instances accurately and (20+14) or 34 test instances incorrectly.

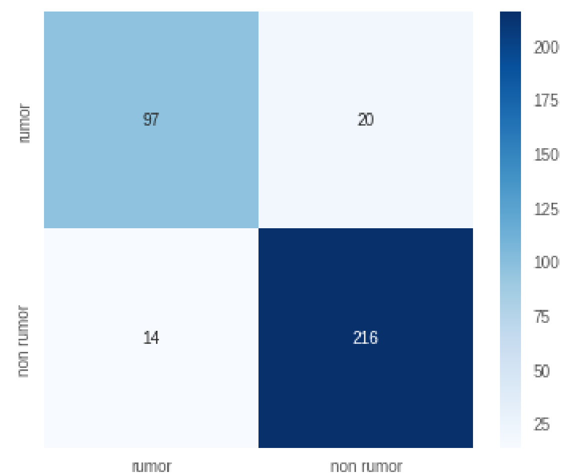

Figure 14: Confusion matrix of BERT

From figure 15 we can see that the confusion matrix of DistilBERT model predicted (93+216) or 309 test instances accurately and (23+15) or 38 test instances incorrectly.

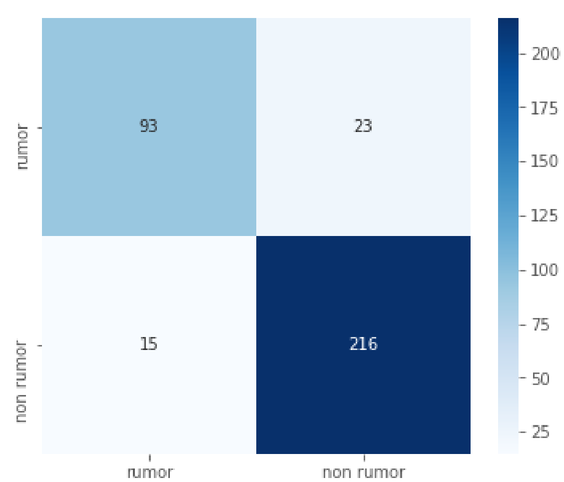

Figure 15: Confusion matrix of SVM

ROC Curve

From figure 16, we found that BERT model has an AUC value of 0.88 that indicates that this is the best performing transformer based deep learning model than DistilBERT model.

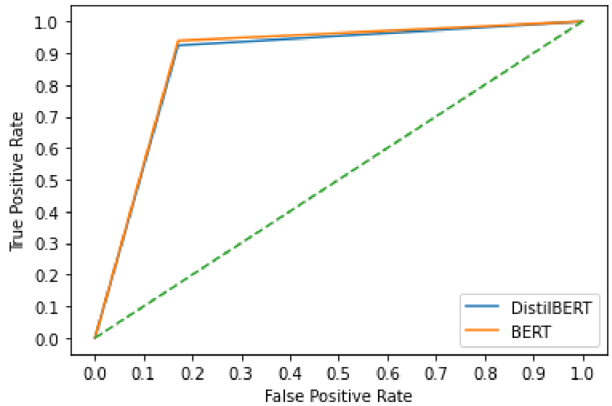

Figure 16: ROC Curve of Deep Learning Models

Fbeta Measure

The table 6 shows that the number of false positive predictions is substantial for all three models.

Table 6: Fbeta Measure of Deep Learning Models

Models	Fbeta(=0.5)	Fbeta(=1)	Fbeta(=2)
BERT	0.9199	0.9270	0.9343
DistilBERT	0.9111	0.9159	0.9208

Because increasing accuracy decreases false positives, we must evaluate the Fbeta(=0.5)-measure from table 7, which prioritizes reducing false positives above minimizing false negatives.

Table 7: False Predictions of Deep Learning models

Models	False Positive	False Negative
BERT	20	14
DistilBERT	23	15

Comparison Analysis

Here in this section, we have done comparison analysis to find the best performing model between the two groups. From figure 17, we can identify that among the machine learning-based models (XGBoost, Random Forest, and SVM), the SVM has successfully achieved the highest accuracy (87.56%). On the other hand, for the deep learning-based transformers, the BERT has achieved a higher accuracy of 90.20% compared to the 89.05% accuracy of the DistilBERT.

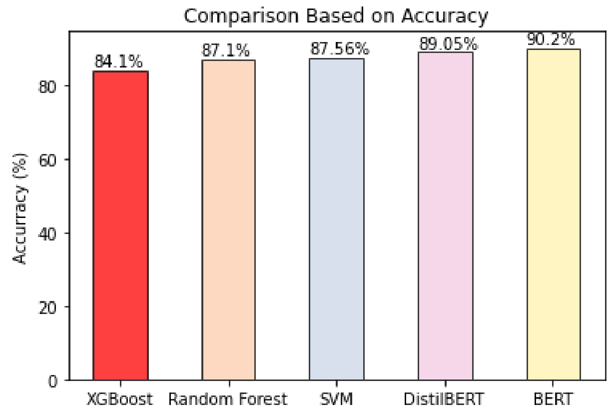

Figure 17: Comparison Analysis of Our Models

So, from this study, we concluded that the deep learning-based transformer models comparatively provide better performance than the traditional machine learning-based models. In addition to this, among all the baseline models(XGBoost, Random Forest, and SVM) and transformers(BERT and DistilBERT), the BERT is successful in attaining the leading accuracy of 90.20%.

Conclusion

Communication and the exchange of information among people have always been considered challenging tasks. But currently due to the technological advancements, the communication barrier has been removed with the help of several low-cost, easy access, fast communication platforms(e.g., Facebook, Twitter, Reddit, Youtube, and Sina Weibo) as well as with wide the usage of hand-held devices (e.g., mobile phones, tablets, and smartwatch). With this huge propagation of information opportunity, at present, we are faced with a new problem of rumor dissemination. The rumors, as well as their associated negative aftereffects, are constantly growing and this has the potential to incite widespread wrath. By recognizing and deleting rumors early in their life cycle, the whole life cycle of rumors can be broken. In this study, we attempted to find a method for distinguishing between rumors and non-rumors. We gathered labeled datasets comprising thousands of social media postings and used machine learning and deep learning algorithms to predict the outcomes and then compared the findings. The BERT, deep learning transformer outperformed the other methods, and we want to use this knowledge in the future by employing other modeling techniques to recognize rumors and non-rumors.

Future Work

Our current work is focused on the binary classification/detection of a post into rumors and non-rumors. Multi-level classification of rumors into several sub-groups (e.g. rumor, misinformation, disinformation, sarcasm, finger-pointing, etc.) can be done for further understanding of the source and factors working behind the propagation of rumors.

References

Ahsan, M.; Kumari, M.; and Sharma, T. 2019. Rumors detection, verification and controlling mechanisms in online social networks: A survey. *Online Social Networks and Media* 14: 100050.

Asghar, M. Z.; Habib, A.; Habib, A.; Khan, A.; Ali, R.; and Khattak, A. 2021. Exploring deep neural networks for rumor detection. *Journal of Ambient Intelligence and Humanized Computing* 12(4): 4315–4333.

Benecke, O.; and DeYoung, S. E. 2019. Anti-vaccine decision-making and measles resurgence in the United States. *Global pediatric health* 6: 2333794X19862949.

Breiman, L. 2001. Random forests. *Machine learning* 45(1): 5–32.

Chakravarthy, S. 2020. Tokenization for natural language processing.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, Y.; Sui, J.; Hu, L.; and Gong, W. 2019. Attention-residual network with CNN for rumor detection. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1121–1130.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Evgeniou, T.; and Pontil, M. 1999. Support vector machines: Theory and applications. In *Advanced Course on Artificial Intelligence*, 249–257. Springer.

Fan, C.; Chen, M.; Wang, X.; Wang, J.; and Huang, B. 2021. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research* 9: 652801.

Guo, H.; Cao, J.; Zhang, Y.; Guo, J.; and Li, J. 2018. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM international conference on information and knowledge management*, 943–951.

Hosni, A. I. E.; Li, K.; and Ahmad, S. 2020. Minimizing rumor influence in multiplex online social networks based on human individual and social behaviors. *Information Sciences* 512: 1458–1480.

Kumar, A.; and Sangwan, S. R. 2019. Rumor detection using machine learning techniques on social media. In *International conference on innovative computing and communications*, 213–221. Springer.

Liu, X.; Nourbakhsh, A.; Li, Q.; Fang, R.; and Shah, S. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1867–1870.

Ma, J.; Gao, W.; Joty, S.; and Wong, K.-F. 2020. An attention-based rumor detection model with tree-structured recursive neural networks. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11(4): 1–28.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K.-F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks .

Ma, J.; Gao, W.; and Wong, K.-F. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*, 3049–3055.

Mikkelson, D. 2021. Fake CNN Report: COVID Vaccine Recipients Become Cannibals. URL https://www.snopes.com/fact-check/covid19-vaccination-cannibalism/.

Pathak, A. R.; Mahajan, A.; Singh, K.; Patil, A.; and Nair, A. 2020. Analysis of techniques for rumor detection in social media. *Procedia Computer Science* 167: 2286–2296.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* .

Santhoshkumar, S.; and Dhinesh Babu, L. 2020. Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks. *Social Network Analysis and Mining* 10(1): 1–17.

Shelke, S.; and Attar, V. 2019. Source detection of rumor in social network–a review. *Online Social Networks and Media* 9: 30–42.

Tafannum, F.; Shopnil, M. N. S.; Salsabil, A.; Ahmed, N.; Alam, M. G. R.; and Reza, M. T. 2021. Demystifying Black-box Learning Models of Rumor Detection from Social Media Posts. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 0358–0364. IEEE.

Veyseh, A. P. B.; Thai, M. T.; Nguyen, T. H.; and Dou, D. 2019. Rumor detection in social networks via deep contextual modeling. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 113–120.

Yang, L.; Li, Z.; and Giua, A. 2020. Containment of rumor spread in complex social networks. *Information Sciences* 506: 113–130.