Chapter 10

Predicting Catastrophic Events Using Machine Learning Models for Natural Language Processing

Muskaan Chopra
Chandigarh College of Engineering and Technology, India

Sunil K. Singh
https://orcid.org/0000-0003-4876-7190
Chandigarh College of Engineering and Technology, India

Kriti Aggarwal
Chandigarh College of Engineering and Technology, India

Anshul Gupta
Chandigarh College of Engineering and Technology, India

ABSTRACT

In recent years, there has been widespread improvement in communication technologies. Social media applications like Twitter have made it much easier for people to send and receive information. A direct application of this can be seen in the cases of disaster prediction and crisis. With people being able to share their observations, they can help spread the message of caution. However, the identification of warnings and analyzing the seriousness of text is not an easy task. Natural language processing (NLP) is one way that can be used to analyze various tweets for the same. Over the years, various NLP models have been developed that are capable of providing high accuracy when it comes to data prediction. In the chapter, the authors will analyze various NLP models like logistic regression, naive bayes, XGBoost, LSTM, and word embedding technologies like GloVe and transformer encoder like BERT for the purpose of predicting disaster warnings from the scrapped tweets. The authors focus on finding the best disaster prediction model that can help in warning people and the government.

DOI: 10.4018/978-1-7998-8413-2.ch010

INTRODUCTION

Disasters and crises have always been dynamic and chaotic by nature (Guha-Sapir et al., 2016). Natural disasters disrupt people's lives in an instant, causing mental, physical, and societal damage. Communication is important in all aspects of disaster management in such a circumstance (Mukkamala et al., 2016). Earlier people used to use traditional media channels in order to communicate disaster warnings. The traditional media included devices such as television, newspaper, and radio channels. However, this method of communication was slow and difficult.

With the development of new technologies, the communication ability of people has also improved and developed tremendously. One of the major contributors to this change is the ever-evolving smartphone and internet technologies. The second reason is the plethora of software and applications available on these devices. Unlike before, now people are just "a few clicks" and one "app" away from being able to send or receive information.

One of the biggest influencers of all has been the social media applications and platforms like FaceBook, Instagram, and Twitter. All these platforms have played a vital role in the prediction, the announcement of disasters. Not only this, but these platforms have also served as a way for people to communicate and share their grief as well as send and receive help. Today, social media is recognized as one of the most popularly used media platforms for disaster management. (Gray et al., 2016, Ranjan Avasthi, 2017).

Social media gives people the freedom to become producers along with being consumers of information. The present paper focuses on Twitter which is one of the most widely used social media networks for disaster prediction and management research. Since the launch of Twitter in 2005, it has become one of the largest microblogging services. (Dhiraj Murthy, 2018).

With new hashtags and keywords trending on Twitter every day, the latest trend is spreading the news updates from around the world. This is something that has proved to be beneficial in the prediction of natural disasters and emergency situations (Ulvi, O et al, 2019) like floods (Vieweg S. et al., 2010, Vieweg S. et al., 2010, Starbird K et l., 2010), earthquakes (Earle, P et al., 2010, Kireyev, K, et al., 2009, Muralidharan S et al), wildfires, and hurricanes (Hughes AL, Palen L, 2009, Hughes AL et al., 2008), etc. The platform has shown great potential in increasing the survival rate during emergency situations caused due to disasters like tornadoes and wildfires. In one such paper (Lindsay B, 2011), authors emphasized the use of Twitter as a multidirectional communication network that can not only aid officials in compiling lists of the disaster-affected people but also help them in contacting the grieved family members. Tweeting has hence provided people a way to announce emergency situations in real-time. People use this channel to express and describe what they see and observe. Programmatically monitoring tweets can hence help in finding hints for upcoming calamities. Many disaster relief organizations and news agencies hence have been aiming to develop technologies and programs related to tweet analysis.

One point to take note of is that it cannot be guaranteed that a particular piece of information found on social media is "Truth" (Zhang, Z., & Gupta, B. B., 2018). Moreover, it is not easy to identify emergency tweets just by analyzing the words. In order to actually declare them as disaster announcements, much in-depth analysis is required. (Iruvanti, G, 2020). A study done in (Kate Starbird et al., 2010) exemplifies the dilemma of conversational microblogging. They focused on the use of Twitter for communicating the disasters where they found high discrepancies in the tweets and retweets.

The majority of the material retweeted during the 2009 Red River disaster in North Dakota was the one that previously existed on Twitter. However, the news that spread included less than 10% of the real tweets. The majority of news floating onto mainstream channels was found to be derivative. These synthetic and fake tweets spread like wildfire in the wake of original tweets.

The present paper hence focuses on developing a Natural Language Processing (NLP) based Disaster Prediction model. This model will help people as well as the government to detect whether a particular tweet is a premonition of disaster or a false warning. The result of this analysis will not only help people to be well aware of what is about to come but will also help in preventing the creation of fake commotion.

In section two of the present papers, the authors discuss the previous research works being done in the field of disaster prediction using Twitter. Section 3 briefs about the NLP techniques used in the paper. This section discusses different processes like data pre-processing and building prediction models. The last section of the paper compares these models based upon their test accuracy to find the most suitable model of disaster prediction.

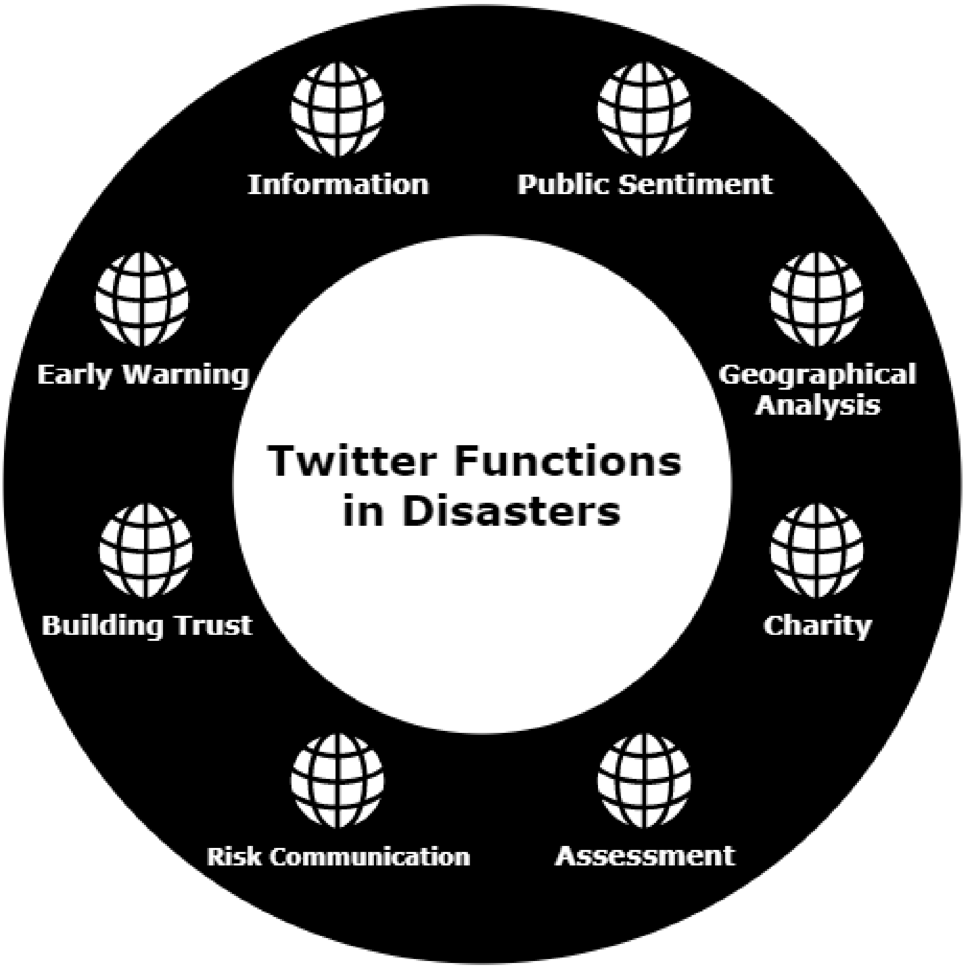
LITERARY WORK & MOTIVATION

Twitter is one of the most popular microblogging sites. The platform receives millions of tweets on a daily basis. According to a survey done in (Srivastava, A et al., 2019), every second twitter disseminates around 6,000 tweets. This amounts to almost 500 million tweets per day and more than 200 billion tweets on an annual basis. The tweets range from informational to advertising and promotional content. However, in the current era of digitalization, where people are the content generators, determining the truth value of the information is very important (Mendoza et al., 2010). Many researchers have noted a surge in the spread of fake news especially during emergency situations like the Covid-19 pandemic (Gupta A et al., 2021, Sahoo, S. R., & Gupta, B. B., 2021).

The research problem in the present paper is motivated by various researchers that show the exponential increase in the frequency of tweets during any disaster period (Lamsal, R., & Kumar, T. V., 2021). Careful monitoring, processing, and analyses of these tweets can help in the identification of important information like a prediction of calamity, call for help, identification of missing people, etc. Discussions on Twitter vary between thousands of different topics. However, some studies suggest that the maximum number of daily users are only interested in viral tweets consisting of hot topics and top news (Murthy, J. S. et al., 2019). The number of Twitter users, as well as the number of tweets, increase daily, yet, the participation of official organizations during emergency situations and crises remains negligible (Latonero, M., & Shklovski, I., 2011).

Further studies and research work suggest that the automatic classification of tweets can play a crucial role in the identification of emergency situations. It can also help official organizations to take well-timed action, thus saving the lives of affected people. Over the years, a number of classification algorithms and solutions have been successfully used to detect and classify natural disaster tweets and messages. Some researches even extended beyond the simple disaster prediction and classification phase by enhancing the results using visualization techniques like mapping. One of the papers (Ian P. Benitez et al., 2018), used a feature vector matrix for the purpose of representing features extracted from Twitter messages. They applied an improved Genetic Algorithm for the extraction of features. Social media data supplied by catastrophe witnesses have been found to be extremely valuable for disaster management and response. To deal with the paucity of labeled data at the start of a target catastrophe, domain adaptation techniques have been utilized. The approaches for designing the models range from conventional machine learning algorithms (Li et al.2017) to newer deep learning-based ones (Li, X., & Caragea, C. 2019). One such paper (Li et al.2017) proposed a self-training type approach. The approach used Naïve Bayes as the base classifier. (David Graf et al., 2018) also designed a cross-domain classifier for different disaster types. Xukun Li and Doina Caragea (Li, Xukun & Caragea, Doina., 2020) surveyed various DRCN approaches for disaster tweet classification. Figure 1 Previously done work in Disaster predictions using Twitter (Seddighi H et al., 2020) shows the collective research topics and work that has been done on various disaster prediction phases using Twitter.

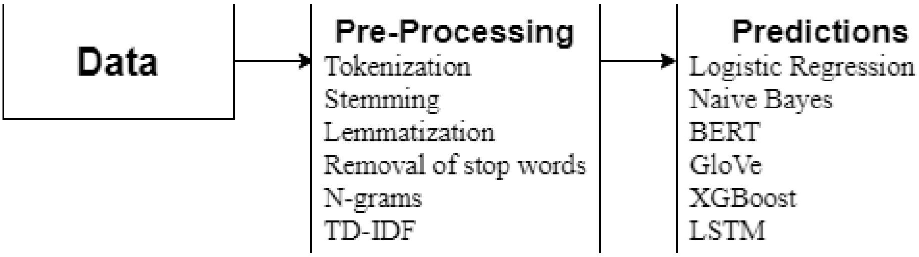
Figure 1. Previously done work in Disaster predictions using Twitter
Source: Seddighi H et al., 2020



METHODOLOGIES

For the prediction and analysis of disasters in Tweets, several procedures were achieved. This process is illustrated in Figure 2 and the steps are further explained in detail in the following sections.

Figure 2. Methodology and workflow



Dataset Description

The data set used in the present paper for the identification of disaster and emergency-related tweets were taken from an online open-source resource. The dataset was originally created by figure-eight, an AI ML-based company, popular for providing high trained datasets. The data set was initially shared on the website ‘Data For Everyone’(Datasets resource center, 2021). One of the issues which the authors faced while collecting Twitter data is filtering tweets. Tweets often contain irrelevant information like advertisements and promotional materials and spamming activities (Sahoo, S. R., & Gupta, B. B. 2020). Various researchers have studied and dealt with this problem like (Wang, A. H., 2010).

Exploratory Data Analysis

Exploratory data analysis is a process for taking data insights and summarizing their main properties. It enables us to make some inferences from the data by visualizing it and exploring its statistical properties (Oleksandr Zaytsev et al., 2017). Visualization of tweets can be seen in Figure 4 and Figure 5 where the text is divided on the basis of location as well as categorized into disastrous and non-disastrous.

EDA can be particularly helpful to get information about the inside structure of a large dataset, detecting missing values, determining the relations among the dataset through visualizations, and selecting particular models for further predictions and analysis. This is shown in Figure 3.

Figure 3. Identification of Missing Values through EDA in the dataset

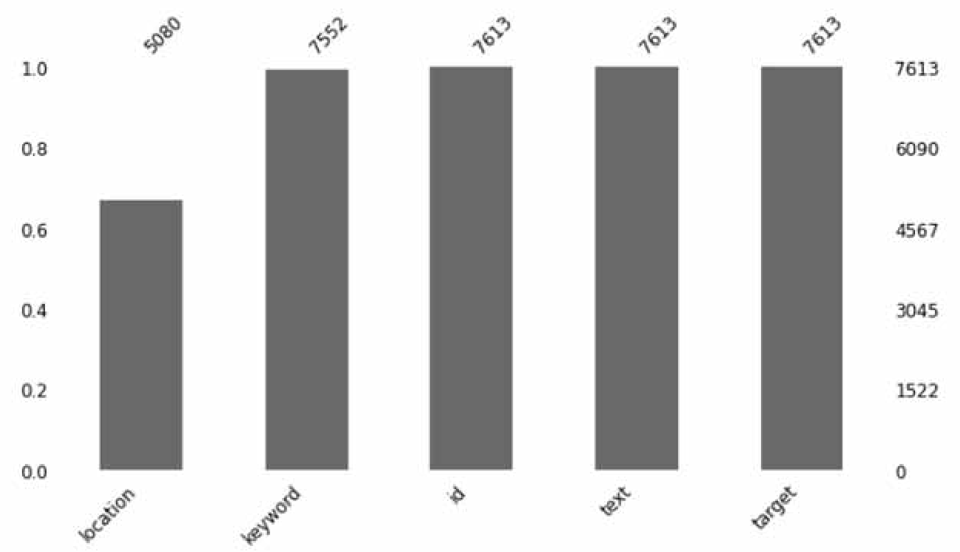


Figure 4. Classification of Tweets by location as Non- Disaster

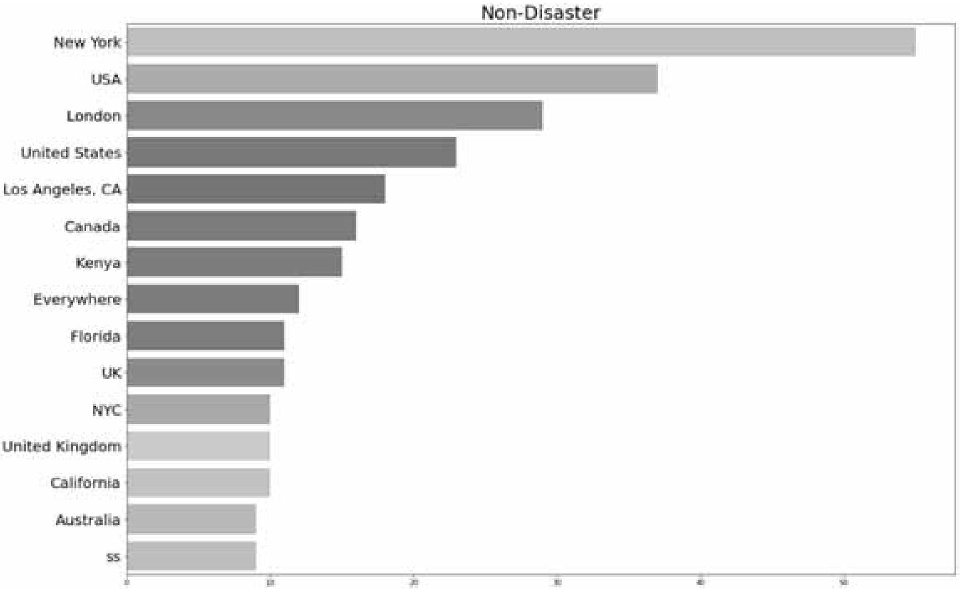
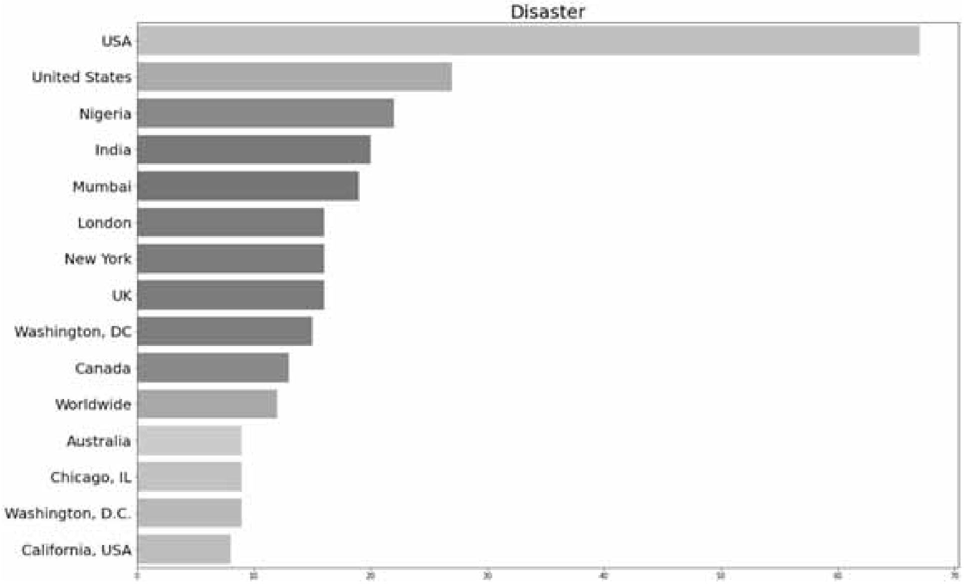


Figure 5. Classification of Tweets by location as Disaster



Data Pre-Processing

A number of processes were carried out for data analysis and preprocessing which included natural language processes like tokenization, stemming, removal of stop words, frequency count, lower case conversions, and advanced textual formatting. Lowercasing is the first step for data preprocessing which helps in the removal of duplicates. This step helps in the calculation of accurate word count. The next step is removing punctuations and stop words so that no extra information is added while treating the textual data. These steps reduce the overall size of the data set considerably. One of the important steps in data preprocessing is tokenization. Tokenization helps in the division of the text into a sequence of words. After this step, stemming and lemmatization are done to further improve data quality. The present paper uses lemmatization over stemming because while stemming removes suffices, lemmatization performs morphological analysis by converting word to root. Through these processes, all the null and repetitive data was removed. These processes were performed to obtain the suitable blend of preprocessing tasks required for better predictions and results. Before the predictions and model application, the data was vectorized. N-grams were used in order to create unigram, bigram, and trigrams which help in capturing language structures. Term Frequency and Inverse Document Frequency were used to generate TF-IDF. TD is the ratio of the word count present in a sentence to the length of the sentence. IDF value indicates the uniqueness of a word, the more the IDF value, the more unique the word is (Analytics Vidhya, 2020).

Figure 6. Most occurring after cleaning and preprocessing the data set

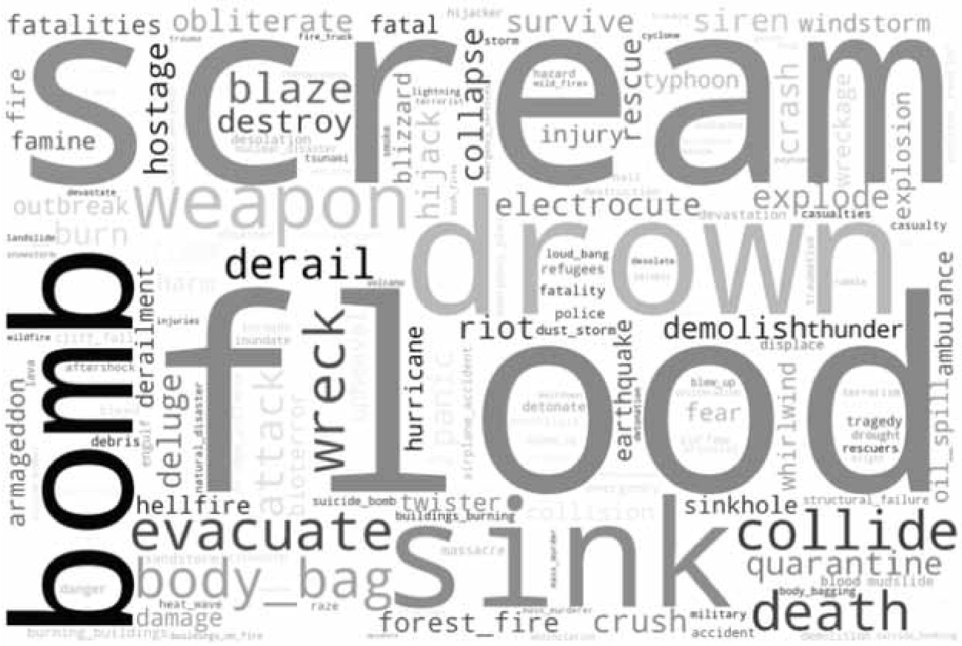


Figure 6 shows the keywords or the most occurring words in the datasets through a word cloud visualization. Furthermore, preprocessed and cleaned data has been used for training the models. The models have been discussed in detail in the next section.

MODELS

Logistic Regression

One of the popular statistical approaches to Machine Learning and NLP is Logistic regression. This technique uses relationships of outputs or independent variables or dependent variables. The result is used to analyze whether the dependent variable is dichotomous, unordered polytomous, or ordered polytomous:

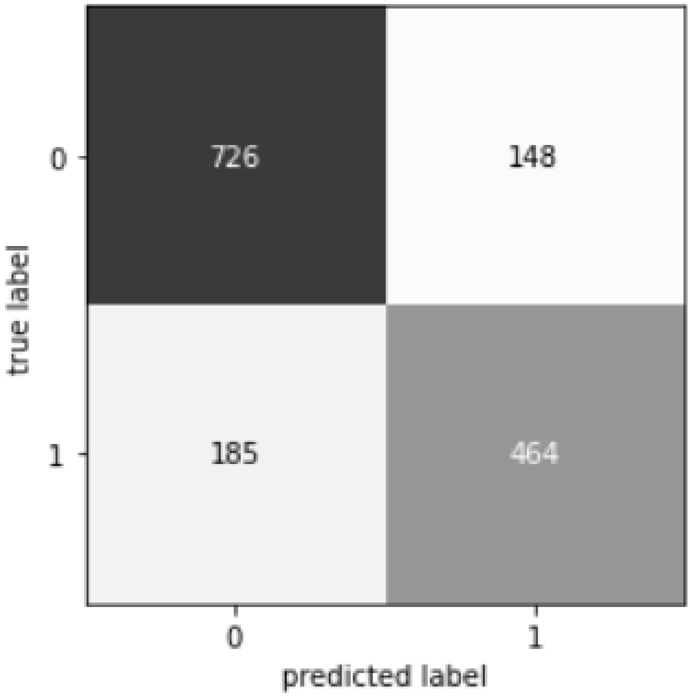
(a) dichotomous:- A dependent variable is dichotomous when it has only two categories. Some examples of this are categorizing data into negative and positive.

(b) unordered polytomous:- An unordered polytomous is a nominal scale variable having three or more classifications. One of the popular examples of this can be seen in the case of the political party identification problem.

(c) ordered polytomous: Unlike unordered polytomous, an ordered polytomous is an ordinal scale variable having three or more classifications. An example of this can be seen in (Salkind, N. J. 2010), where the researchers have used it to complete the level of education.

In the present paper, the dataset is categorized into only two categories. The first category is the real disaster represented by 1 and the second one is the fake disaster represented by 0, it falls under dichotomous. The model is used over the preprocessed data set to predict the results as shown in Figure 7.

Figure 7. Confusion Matrix for Logistic Regression

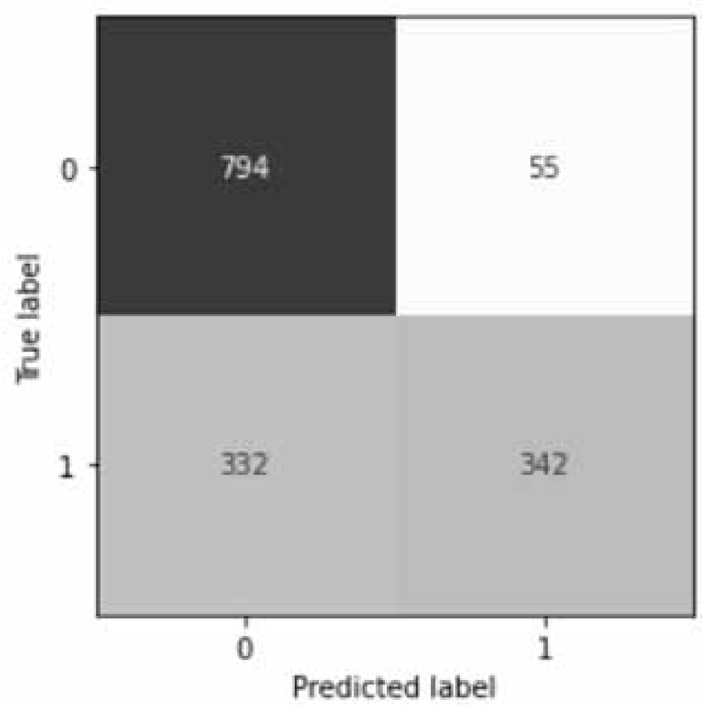


The confusion matrix in figure 7 gives the description of the classifier. The model gives an accuracy of 0.781 on the data set.

XGBoost

The second model is the XGBoost which is another version of the gradient boosted decision tree classifier. In this model, the trees are built sequentially. The goal of each succeeding tree is to decrease the mistakes of the preceding tree. These succeeding trees are known as basic or weak learners. Each of these weak learners offers some crucial information for prediction, allowing the boosting approach to successfully combine these weak learners to generate a strong learner. The strength of XGBoost resides in its scalability, which enables rapid learning via parallel and distributed computation and economical memory consumption. The confusion matrix in figure 8 gives the description of the classifier. The model gives an accuracy of 0.745 on the data set.

Figure 8. Confusion Matrix for XGBoost



Naive Bayes

Naive Bayes uses prior information to compute a posterior probability which is represented as a probability distribution. Probability distribution indicates the likelihood of a particular instance belonging to a specific class. This algorithm is said to be "naive" as it makes two important assumptions about the environment which are defined as follows:

1. Each variable is statistically independent of the others
2. Each characteristic is equally important.

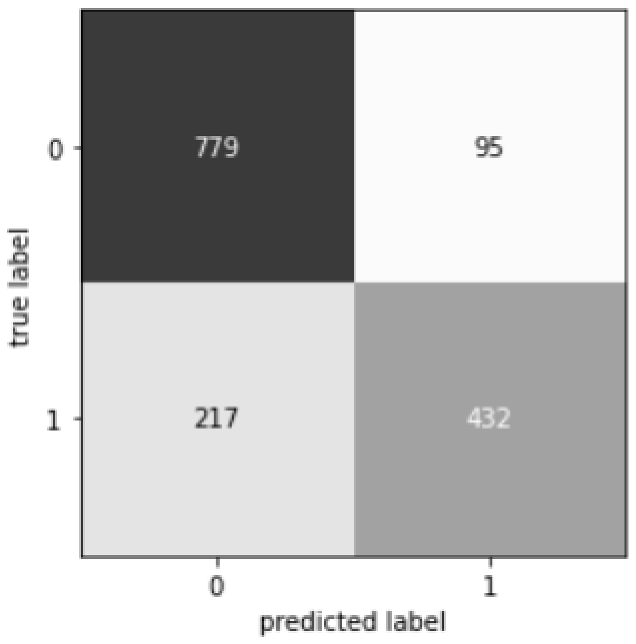
Although these two criteria are extremely unusual in reality, these simplifying assumptions help in deducing the likelihood of a certain occurrence. This is done using the basic idea of Bayesian conditional probabilities and decision making.

The naive Bayes classifier has the benefit of using a little quantity of training data to estimate the parameters. The drawback of NB is the class conditional independence assumption, which means that NB loses accuracy when there are relationships among variables. Dependencies between variables cannot be represented by naive Bayesian Classifiers, but they may be dealt with using Bayesian Belief Networks (Saad, Motaz, 2010). There are two different ways in which naive Bayes can be represented: Naive Bayes Complement and Naive Bayes Multinomial.

Naive Bayes Complement

Naive Bayes complement is highly preferred with small but imbalanced datasets. It is used to calculate a posterior probability with the help of prior knowledge. This is represented by a probability distribution which calculates the probability of a specific instance belonging to all classes (Saad, Motaz, 2010). The confusion matrix in figure 9 gives the description of the classifier. The model gives an accuracy of 0.795 on the data set.

Figure 9. Confusion Matrix for Naive Bayes Complement



Naive Bayes Multinomial

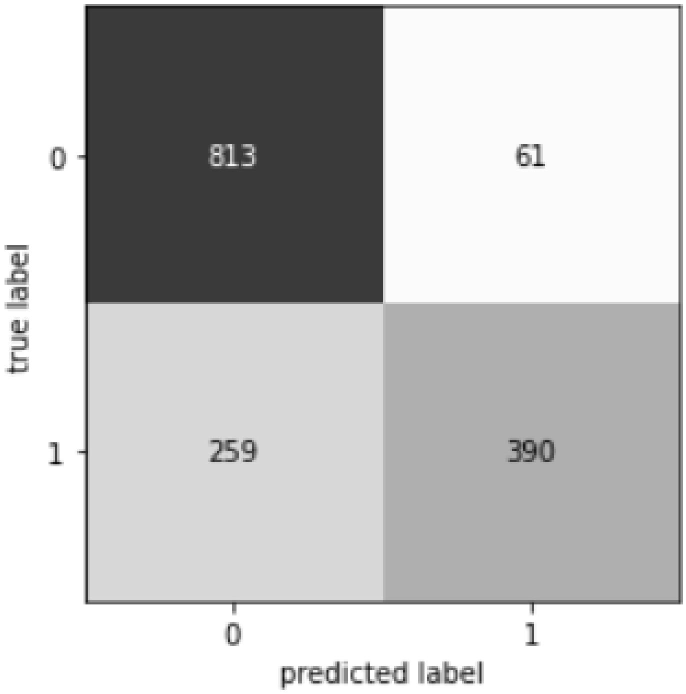
This method employs basic, heuristic solutions to some of the issues associated with naive Bayes classifiers. The technique tackles both fundamental difficulties as well as those that occur as a result of the text not being created using a multinomial model (Saad, Motaz, 2010). The confusion matrix in figure 10 gives the description of the classifier. The model gives an accuracy of 0.789 on the data set.

LSTM

Long Short-Term Memory Networks (LSTMs) are Recurrent Neural Network (RNN) versions that tackle RNN gradient vanishing/exploding issues. LSTMs are intended to capture long-distance relationships within texts. Each LSTM unit has three gates that govern which information to remember, forget, and pass to the next phase. LSTMs maintain lengthy dependencies between words and hold the contextual meaning of each word based on the surrounding information. However, they only pay attention to one direction of information, which is the past. Bi-LSTMs, on the other hand, concentrate on the input's past and future orientations (O'Keefe, Simon & Alrashdi, Mohammed, 2018). This technique allows the network to collect more information than before. This is done by concatenating hidden representations, at each token location, from each direction. This method gives an accuracy of 0.719.

GloVe

Figure 10. Confusion Matrix for Naive Bayes Multinomial



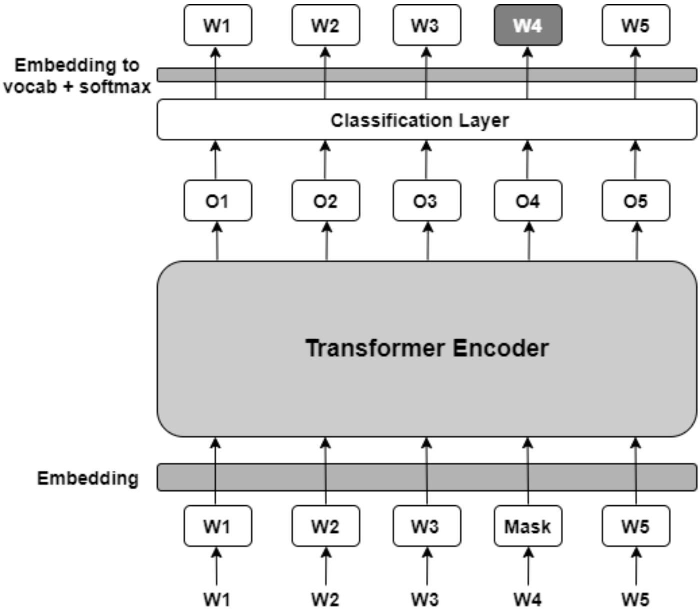
The application of word embedding has piqued the curiosity of many NLP researchers in recent years. Word embedding is a class of feature learning approaches or language models in which texts (words or phrases) are mapped to real-world vectors of numbers. The primary objective of word embedding is to develop expressive and efficient text representations in which related words or phrases have representations capable of conveying their semantic meaning. (Naili et al., 2017).GloVe embedding is a well-known universal pre-trained word embedding that the authors in (Pennington et al., 2014) produced. This embedding has been shown to have an important impact in the improvement of several NLP tasks (Pennington et al., 2014). GloVe embedding is a freely accessible 100-dimensional embedding that has been trained on 6 billion words from online text and Wikipedia and is comparable to tweets (O'Keefe, Simon & Alrashdi, Mohammed, 2018). This method gives an accuracy of 0.81.

BERT

Because of its capacity to capture contextual word embeddings, BERT and its variations have been widely utilised as building blocks for a wide range of applications (Song, G.; Huang, D, 2021). Transformers are altering the entire path of NLP in order to get SOTA (State of the Art) outcomes.

BERT is an acronym that stands for Bidirectional Encoder Representations from Transformers. BERT is a model that is "deeply bidirectional." The use of the term bidirectional, here, indicates the ability of BERT to learn information from both the left and right sides of a token's context, during the training phase. A model's bidirectionality is critical for properly comprehending the meaning of a language. This approach gives the highest accuracy of 0.846 on the data set.

Figure 11. High-level description of the Transformer encoder



CONCLUSION

The field of Tweet analysis and predictions has witnessed some impressive advancements in recent years. Initially, the extracted and scraped tweets were analyzed and visualized. Things have come a long way, and advanced natural language processing models are being used instead. The results have also improved steadily and are tending more and more towards realism.

In this review, the authors have studied various models proposed for performing the task of natural language processing which include Word Embeddings, Logistic Regression, Naive Bayes, BERT. Table I compares the accuracy of the above-mentioned Natural Language Processing Models trained on the Disaster Tweet Dataset.

Table 1. Test Accuracy of Different Models for Disaster Prediction using Tweets

Model Used for Prediction	LSTM	XGBoost & Decision Trees	Logistic Regression	Naive Bayes Multinomial	Naive Bayes Complement	GloVe	Transformer / BERT
Test Accuracy	0.719	0.745	0.781	0.789	0.795	0.813	0.846

Figure 12. Test Accuracy of Different Models for Disaster Prediction using Tweets

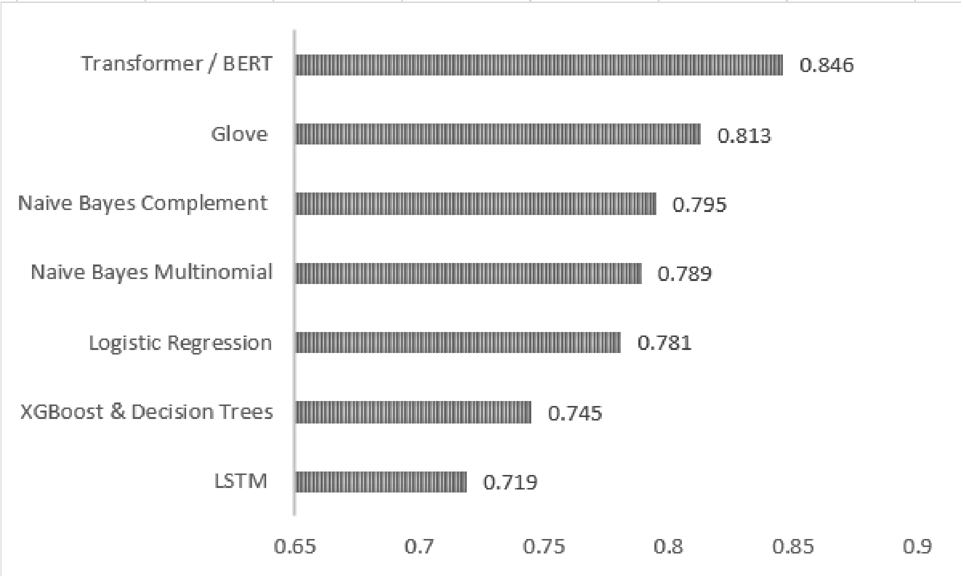


Figure. 12 gives a pictorial representation of the data which is provided in tables I. From Figure. 12, the Transformer/BERT has generated the most impressive results as compared to the other models. On the basis of the values, the authors also conclude that the results of the Naive Bayes are only slightly better than the Logistic Regression. Similarly, XG Boost and LSTM have also shown similar results, but XG Boost does perform better on the dataset due to its effectiveness in dealing with large datasets.

Notable developments in the future will involve improving the training stability of these models and increasing the accuracy of predictions in real-time also.

REFERENCES

Al-Qurishi, M., Rahman, S. M. M., Alamri, A., & Mostafa, M. A., & Al-Rubaian. (2018). SybilTrap: A graph-based semi-supervised Sybil defense scheme for online social networks. *Concurrency and Computation*, *30*(5), e4276.

Avasthi, R. (2017). Social Media and Disasters: A Literature Review. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(10). doi:10.1016/j.jaac.2017.07.317

Benitez, Sison, & Medina. (2018). Implementation of GA-Based Feature Selection in the Classification and Mapping of Disaster-Related Tweets. In *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval (NLPIR 2018).* Association for Computing Machinery. doi:10.1145/3278293.3278297

Bouarara, H. A. (2021). Recurrent Neural Network (RNN) to Analyse Mental Behaviour in Social Media. *International Journal of Software Science and Computational Intelligence*, *13*(3), 1–11.

Candon, P. (2019). Twitter: Social communication in the Twitter era. *New Media & Society*, *21*. doi:10.1177/1461444819831987

Chaudhary, P., Gupta, B. B., & Yamaguchi, S. (2016, October). XSS detection with automatic view isolation on online social network. In *2016 IEEE 5th Global Conference on Consumer Electronics* (pp. 1-5). IEEE.

Chen, T. Y., Chen, Y. M., & Tsai, M. C. (2020). A Status Property Classifier of Social Media User's Personality for Customer-Oriented Intelligent Marketing Systems: Intelligent-Based Marketing Activities. *International Journal on Semantic Web and Information Systems*, *16*(1), 25–46.

Datasets Resource Center. (2021, March 11). *Appen*. https://appen.com/open-source-datasets/

Disasters on social media - dataset by crowdflower. (2016, November 21). https://data.world/crowdflower/disasters-on-social-media/access.

Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010, March). OMG Earthquake! Can Twitter improve earthquake response? *Seismological Research Letters*, *81*(2), 246–251.

Graf, D., Retschitzegger, W., Schwinger, W., Pröll, B., & Elisa-beth, K. (2018). Cross-domain informativeness classification for disaster situations. *Proceedings of the 10th international conference on management of digital ecosystems*, 183–190.

Gray, B., Weal, M., & Martin, D. (2016). *Social media and disasters: A new conceptual framework*. Academic Press.

Guha-Sapir, D., Vos, F., Below, R., & Ponserre, S. (2016). Annual Disaster Statistical Review 2016: The Numbers and Trends. Brussels: Centre for Research on the Epidemiology of Disasters (CRED).

Gupta, A. (in press). An Exploratory Analysis on the Unfold of Fake News During COVID-19 Pandemic. *Smart Systems: Innovations in Computing*.

Gupta, S., Gupta, B. B., & Chaudhary, P. (2018). Hunting for DOM-Based XSS vulnerabilities in mobile cloud-based online social network. *Future Generation Computer Systems*, *79*, 319–336.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *ISCRAM Conference*.

Hughes, A. L., Palen, L., Sutton, J., Liu, S., & Vieweg, S. (2008). *"Site-seeing" in disaster: an examination of on-line social convergence'*. Paper presented at the ISCRAM (Information Systems for Crisis Response and Management) Conference, Washington, DC.

Iruvanti, G. (2020, August 1). *Real or NOT? NLP with DISASTER Tweets (classification using google bert).* Medium. https://levelup.gitconnected.com/real-or-not-nlp-with-disaster-tweets-classification-using-google-bert-76d2702807b4

Kireyev, K., Palen, L., & Anderson, K. (2009). Applications of topics models to the analysis of disaster-related Twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond* (*Vol. 1*). Academic Press.

Lamsal, R., & Kumar, T. V. (2021). Twitter-Based Disaster Response Using Recurrent Nets. *International Journal of Sociotechnology and Knowledge Development (IJSKD), 13*(3), 133-150. doi:10.4018/IJSKD.2021070108

Latonero, M., & Shklovski, I. (2011). Emergency Management, Twitter, and Social Media Evangelism. *International Journal of Information Systems for Crisis Response and Management*, *3*(4), 1–16. https://doi.org/10.4018/jiscrm.2011100101

Li, X., & Caragea, C. (2019). *Identifying Disaster Damage Images Using a Domain Adaptation Approach*. ISCRAM.

Li & Caragea. (2020). *Domain Adaptation with Reconstruction for Disaster Tweet Classification.* . doi:10.1145/3397271.3401242

Li, Caragea, Caragea, & Herndon. (2017). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management, 26*. doi:10.1111/1468-5973.12194

Lindsay B. (2011). *Social media and disasters: Current uses, future options and policy considerations.* Washington, DC: Congressional Research Service CRS Report for Congress, Analyst in American National Government.

Mendoza, M., Poblete, B., & Castillo, C. (2010). *Twitter under Crisis: Can we trust what we RT?* Paper presented at First Workshop on Social Media Analytics, Washington, DC.

Mukkamala, A., & Beck, R. (2016). *Enhancing Disaster Management Through Social Media Analytics To Develop Situation Awareness: What Can Be Learned From Twitter Messages About Hurricane Sandy?* Academic Press.

Muralidharan, S., Rasmussen, L., Patterson, D., & Shin, J. (2011). Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts. *Public Relations Review*, *37*(2), 175–177.

Murthy, J. S., G.M., S., & K.G., S. (2019). A Real-Time Twitter Trend Analysis and Visualization Framework. *International Journal on Semantic Web and Information Systems (IJSWIS), 15*(2), 1-21. doi:10.4018/IJSWIS.2019040101

Naili, Habacha, & Ben Ghezala. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science, 112*, 340-349. . doi:10.1016/j.procs.2017.08.009

Noor, S., Guo, Y., Shah, S. H. H., Nawaz, M. S., & Butt, A. S. (2020). Research synthesis and thematic analysis of twitter through bibliometric analysis. *International Journal on Semantic Web and Information Systems*, *16*(3), 88–109.

O'Keefe. (2018). Deep Learning and Word Embeddings for Tweet Classification for Crisis Response. Academic Press.

Pennington, Socher, & Manning. (2014). Glove: Global Vectors for Word Representation. *EMNLP, 14*, 1532-1543. . doi:10.3115/v1/D14-1162

Saad, M. (2010). *The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification.* doi:10.13140/2.1.4677.2164

Sahoo, S. R., & Gupta, B. B. (2020). Classification of spammer and nonspammer content in online social network using genetic algorithm-based feature selection. *Enterprise Information Systems*, *14*(5), 710–736.

Sahoo, S. R., & Gupta, B. B. (2021). Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, *100*, 106983.

Salkind, N. J. (2010). *Encyclopedia of research design* (Vol. 1-0). SAGE Publications, Inc., doi:10.4135/9781412961288

Seddighi, H., Salmani, I., & Seddighi, S. (2020). Saving Lives and Changing Minds with Twitter in Disasters and Pandemics: A Literature Review. *Journalism and Media.*, *1*(1), 59–77. https://doi.org/10.3390/journalmedia1010005

Sharma, Y., Bhargava, R., & Tadikonda, B. V. (2021). Named Entity Recognition for Code Mixed Social Media Sentences. *International Journal of Software Science and Computational Intelligence*, *13*(2), 23–36.

Song, G., & Huang, D. (2021). A Sentiment-Aware Contextual Model for Real-Time Disaster Prediction Using Twitter Data. *Future Internet, 13*, 163. doi:10.3390/fi13070163

Srivastava, A., Singh, V., & Drall, G. S. (2019). Sentiment Analysis of Twitter Data: A Hybrid Approach. *International Journal of Healthcare Information Systems and Informatics*, *14*(2), 1–16. doi:10.4018/IJHISI.2019040101

Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010). Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work (CSCW '10).* Association for Computing Machinery. doi:10.1145/1718918.1718965

Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010). Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *CSCW '10: Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM.

Ultimate guide to deal with text data. (2020, December 23). *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/

Ulvi, O., Lippincott, N., Khan, M. H., Mehal, P., Bass, M., Lambert, K., Lentz, E., & Haque, U. (2019, December 10). The role of social and mainstream media during storms. *Journal of Public Health and Emergency*. https://jphe.amegroups.com/article/view/5543/html

Vieweg, S. (2010). Microblogged contributions to the emergency arena: Discovery, interpretation, and implications. Computer Supported Collaborative Work.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*. ACM.

Wang, A. (2010). *Don't Follow Me - Spam Detection in Twitter*. doi:10.7312/wang15140-003

Wang, H., Li, Z., Li, Y., Gupta, B. B., & Choi, C. (2020). Visual saliency guided complex image retrieval. *Pattern Recognition Letters*, *130*, 64–72.

Yen, S., Moh, M., & Moh, T. S. (2021). Detecting Compromised Social Network Accounts Using Deep Learning for Behavior and Text Analyses. *International Journal of Cloud Applications and Computing*, *11*(2), 97–109.

Zaytsev, O., Papoulias, N., & Stinckwich, S. (2017). Towards Exploratory Data Analysis for Pharo. In *Proceedings of the 12th edition of the International Workshop on Smalltalk Technologies (IWST '17).* Association for Computing Machinery. doi:10.1145/3139903.3139918

Zhang, L., Zhang, Z., & Zhao, T. (2021). A Novel Spatio-Temporal Access Control Model for Online Social Networks and Visual Verification. *International Journal of Cloud Applications and Computing*, *11*(2), 17–31.

Zhang, Z., & Gupta, B. B. (2018). Social media security and trustworthiness: Overview and new direction. *Future Generation Computer Systems*, *86*, 914–925.

Zhang, Z., Jing, J., & Wang, X. (2020). A crowdsourcing method for online social networks security assessment based on human-centric computing. *Human-centric Computing and Information Sciences*, *10*, 1–19.

Zhang, Z., Sun, R., Zhao, C., Wang, J., & Chang, C. K. (2017). CyVOD: A novel trinity multimedia social network scheme. *Multimedia Tools and Applications*, *76*(18), 18513–18529.