Prediction Of Real Disaster Tweets Using Natural Language Processing (NLP)

Paujebaselvinraj S,
Computer Science and Engineering, Karunya Institute of Technology and Sciences,
Coimbatore, India – 641114.
pauljebaselvinraj@karunya.edu.in

Suriya Prakash A,
Computer Science and Engineering, Karunya Institute of Technology and Sciences,
Coimbatore, India – 641114.
asuriyaprakash@karunya.edu.in

Sam Jonathan C,
Computer Science and Engineering, Karunya Institute of Technology and Sciences,
Coimbatore, India – 641114.
samjonthan@karunya.edu.in

Evangelin Sonia SV,
Computer Science and Engineering, Karunya Institute of Technology and Sciences,
Coimbatore, India – 641114.
evangelinsonia@karunya.edu

Mano Ranjitham R,
Computer Science and Engineering, Karunya Institute of Technology and Sciences,
Coimbatore, India – 641114.
manoranjitham@karunya.edu

Philip Paul Arunodhayam T,
Computer Science and Engineering, Karunya Institute of Technology and Sciences,
Coimbatore, India – 641114.
tphilippaul@karunya.edu.in

Abstract—**Real-time information during a disaster is often shared on social media sites like Twitter. Disaster relief and response teams can immediately prioritize their efforts with this information. Twitter and other social media networks generate tremendous quantities of data that is unstructured, which text analysis and artificial intelligence (AI) algorithms may sift through to uncover significant terms and keywords linked with disasters. The algorithm may have trouble distinguishing between literal references to disasters and the more common use of those keywords in metaphor, leading to widespread incorrect labeling of tweets. Therefore, the purpose of this research is to use classification models and Natural Language Processing to identify genuine and fabricated disaster tweets (NLP). The dataset includes tweets about both real and fictitious disasters, and it was obtained from the Kaggle website. Additionally, the RStudio software for exploratory data analysis (EDA), highlight selections, and data cleansing before data exhibiting was used to test out two different approaches to testing split. We also developed four classifiers (SVM, KNN, Naive Bayes, and XGBoost) to help with this. When compared to KNN and Naive Bayes, SVM, and XGBoost performed better, with 80% and 78% accuracy, respectively. An 80/20 split, based on the whole dataset rather than a random selection, yielded the highest rates of accuracy. Both KNN (99% accuracy) and Naive Bayes (65% accuracy) performed poorly. Tweets about disasters can be analyzed using natural language processing (NLP) to determine which are reliable. Tools like machine learning, sentiment analysis, and text classification can help with this. Following its training on a corpus of annotated tweets, a classifier can forecast the categorization of unclassified tweets (tweets that have been manually annotated to denote if they pertain to a genuine catastrophe). Additionally, through sentiment analysis, tweets that convey intense urgency, fear, or other emotions frequently linked to real disasters can be discerned. Other factors, such as geographical location and the presence of specific keywords, can also be used to make predictions.**

Keywords—Twitter, Tweet, Artificial Intelligence, Natural Language Processing, Naïve Bayes, AI, NLP, EDA, SVM, KNN, XGBoost

1. INTRODUCTION

1.1 DISASTER TWEETS

The widespread rise in internet users' propensity to tweet has opened a plethora of possibilities for analyzing tweets. Users are tweeting about an infinite variety of topics, the vast majority of which are related to politics, entertainment, sports, and technology. Finding tweets about natural and man-made disasters in the massive amount of data posted every day on Twitter is an interesting challenge. Rescue and relief organizations must get aid to those in need as soon as possible after a natural or man-made disaster. A "disaster tweet" is a social media post relating to a crisis, emergency, or natural disaster. During emergencies, they are typically posted on social media sites like Twitter to alert the public and provide updates. During a disaster, when normal means of communication are likely to be overwhelmed, people can turn to social media for information. By keeping an eye on Twitter during a disaster, first responders can better coordinate rescue efforts, assess the situation on the ground, and pinpoint the most urgent needs. However, remember that not all tweets during a disaster can be trusted. Rescue workers and government agencies have a difficult time doing their jobs because of factors such as a lack of accurate victim location data, an excessive volume of calls for help, and the necessity of prioritizing rescue efforts based on the most pressing needs of victims. Unfortunately, erroneous information shared via social media can lead to chaos and even eyeleteer when taken to its logical extreme. Therefore, before doing anything, make sure the information you have is accurate. In these situations, the response time is prolonged by the lack of

relevant data. During a crisis, timely and accurate information can save lives, and tweets about disasters can play a key role in getting that information to those who need it the most. It has been discovered that during times of crisis, social media sites like Twitter and Facebook see a dramatic increase in the volume of user-generated content. People frequently use these channels to update their status, report casualties and infrastructure damage, provide information about the injured and appeal for aid. In the event of a disaster, the user-generated content produced by social networking sites can be used to coordinate rescue efforts and give people a greater understanding of the situation. Numerous news articles have attested to the significance of social media in aiding in disaster relief, locating aid, and possibly saving lives. One woman was rescued during Hurricane Harvey after tweeting for help when the emergency contact number was out of service. Twenty-eight percent of people surveyed said they would do the same if they were in a similar situation. In the wake of recent disasters, it has become increasingly common for text tweets to be accompanied by related images or videos, thanks to advances in mobile technology. The government and humanitarian organizations can better assess the situation in the aftermath of a disaster when they have access to information from a variety of sources, all in one place. These tweets are a lifeline in times of crisis, but it can be difficult to sift through the noise to find the truly useful ones. Despite the critical information about casualties, missing persons, those who have been located, and damaged buildings and transportation systems, these posts also include several consolation messages and acknowledgment to various organizations helping them. The sheer volume of tweets makes it difficult for humanitarian organizations to manually go through them all and determine which rescue and relief efforts should be prioritized. As a result, there is an urgent need to design a smart system that can sort tweets into various forms of humanitarian assistance. Due to their short length (280 characters maximum) and the frequent use of abbreviations and typos, automatic tweet classification presents significant challenges. Consequently, it is challenging for machine learning classifiers to make sense of them out of context.

1.2 MACHINE LEARNING

Learning by machines is the focus of artificial intelligence (AI) computers to improve themselves through trial and error rather than manual instruction. In the field of machine learning, computers use their analytical prowess to draw conclusions or make decisions based on the data they have analyzed. Supervised learning, unsupervised learning, and reinforcement learning are all types of machine learning. By training on labeled data, a supervised learning model learns to recognize relationships between input and output data. In contrast, unsupervised learning involves training a model on unlabelled data without human intervention to detect patterns and relationships. It is possible to train a model to make decisions depending on feedback from its immediate surroundings. what reinforcement learning is all about maximizing reward. Machine learning has many real-world applications, including advanced areas of research such as recommendation systems, image and speech recognition, natural language processing, and fraud detection. It is being used more and more in industries like transportation, finance, and healthcare to enhance decision-making processes and optimize operations.

1.3 TEXT CLASSIFICATION

Text classification, also known as text categorization, is a type of text labeling performed as part of natural language processing (NLP). Categories can be anything that needs to be identified in the text, including topics, emotions, and intentions. A wide variety of machine learning models, including decision trees, Naive Bayes, logistic regression, support vector machines (SVM), and deep learning models like recurrent neural networks (RNN) and Convolutional neural networks (CNN), can be used for text classification. These algorithms perform analyses on the text data and automatically recognize patterns that define one category over another. The process of labeling texts involves a few different phases. Initial steps in text analytics involve pre-processing the data to remove outliers and clean up the text by removing things like punctuation and stop words. The next step involves the extraction of text data features like word frequencies, n-grams, and TF-IDF values. The extracted features are then used to train a machine-learning model to predict a document's classification. Text classification has various applications, including sentiment analysis, preventing spam, topic modeling, and content categorization…t has many applications, including in business, medicine, economics, and the public sphere. To better organize search results and product recommendations, for instance, text classification can be used in online retail to categorize products according to their descriptions and reviews.

1.4 PREPROCESSING OF TEXT

In natural language processing, text pre-processing is the process of cleaning and morphing textual data information into a format that is ideally adapted for analysis and modeling (NLP). The goal of text pre-processing is to convert the raw text data into a structured representation that can be understood by machine learning algorithms, thereby reducing the amount of noise and redundancy in the data. Pre-processing techniques can vary from job to job and application to application. Text pre-processing can

improve NLP models' accuracy by removing noise and irrelevant information from the data. It can help boost productivity by cutting down on the time spent processing data. Since excessive pre-processing can reduce model performance and lead to the loss of valuable information, finding a happy medium between the two is crucial.

2. LITERATURE REVIEW

2.1 EMOTIONAL CONTENT ANALYSIS OF TWEETS

With the advent of the Internet, people now have a new outlet for sharing their emotions. Because it's a data-rich platform where people can view the opinions of others who have been sorted into discrete groups, it's also developing as an independent critical calculating direction. This paper contributes to the field of sentiment analysis for the classification of customer reviews by analyzing the prevalence of tweets containing highly unstructured opinions that are either positive or negative, or neutral. Meaningful adjectives were extracted from the dataset after it was pre-processed, a feature vector list was selected, machine learning-based classification algorithms were applied, and the dataset was analyzed extraction of synonyms and similarity for the content feature is achieved via a Semantic Orientation-based WordNet and the combination of Naive Bayes, Maximum entropy, and Support Vector Machines (SVM). The classification system was then tested for its recall, precision, and accuracy.

2.2 TWITTER ANALYSIS, CLASSIFICATION INTO POSITIVE, NEGATIVE, AND NEUTRAL EMOTIONS

Because we live in the era of big data, people are constantly bombarded with information. Back then, large social media platforms like Facebook, Twitter, and Instagram handled and stored a lot of information. No type of information cannot benefit greatly from being shared via vast online social networks. It's never been simpler to spread an idea or body of knowledge to a wide audience. This, of course, increases the reliability of the data and immediately raises awareness of the issue of widespread evaluation. Consequently, many organizations have found that social media emotion detection, such as labeling a tweet as positive, negative, or neutral, is a useful tool. The primary objective of this research is to categorize three (3) annotated Twitter datasets into groups representing neutral, positive, and negative perspectives. Oversampling, unigram features, and other features are investigated on the dataset's effects on the overall and class-based accuracy ratios. The second dataset's experiments are back to normal. Overall accuracy in the dataset-1 experiments was 88%, which is better than state-of-the-art. It has been shown that the overall accuracy and the class-based accuracy balance are significantly affected by Unigram characteristics.

2.3 REACTIVE HOTLINE POWERED BY SOCIAL MEDIA DATA AND AI

Rapid and extensive onsite data collection in real-time is essential for flood management. Social media is a relatively new data source that can be used to increase flood preparedness awareness due to its unique ability to disseminate information in the form of real-time, rich data in the form of texts and photos. This research demonstrates the usefulness of social media data for monitoring flood phase transition and locating emergency incidents when combined with other information and processed using Artificial Intelligence (AI) techniques. Using social media data, the author trains a computer vision model to classify images into four distinct phases (preparation, impact, response, and recovery) that can be used to trace the evolution of a disaster. To pinpoint the exact location of critical incidents, the author employs a natural language processing (NLP) model trained with deep learning. The recognized locations' coordinates are assigned by checking a dedicated local gazetteer for the disaster-stricken region, which was quickly compiled using data from the GeoNames gazetteer and the United States Census. The author combined image and text analysis to sift through tweets for the most pertinent details of an incident by focusing on those that feature images from the "Impact" category and precise coordinates. To supplement the automatic data processing, the author performed a manual examination and find that it can further strengthen the AI-processed results to facilitate holistic situational awareness and to set up a passive hotline to inform rescue and search operations. Hurricane Harvey's flooding in the Houston area was used as an example in the developed framework. Figure 1 represents their workflow.

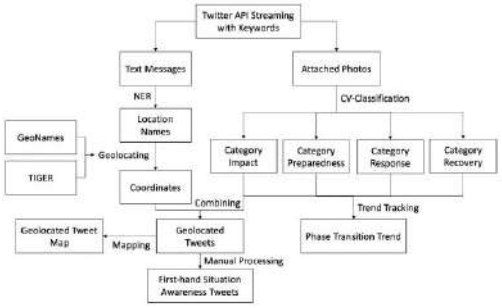

Fig 1: The Social Media Flowchart: Two-Phase Processing

2.3 DISASTER TWEETS AND NATURAL LANGUAGE PROCESSING

After traumatic events like natural disasters, social media is a better way to learn about trends, opinions, and feelings. This paper introduces a model that can identify a real-deal disaster tweet, monitor tweets with a specific hashtag, and glean helpful details from tweets about disasters, such as their location and keywords. This data can be used by the government and non-profits to organize relief efforts like volunteerism, evacuation plans, and donations. analyzing data from various disasters and keeping an eye on Twitter. Using this analysis, we can pinpoint the precise locations of the worst tragedies and most heart-breaking tweets posted by Twitter users. a custom API implemented with this information. This application programming interface provided data on the frequency and location of catastrophic events. The API was used to develop a web-based application. It pinpoints the user's location and provides disaster statistics, such as the severity of disasters in the user's immediate area. The next step is for customers to enter their contact information so they can receive this notification via SMS or email. The primary purpose of this work is to analyze tweets sent during the tragic event to identify the most relevant one.

2.4 TWITTER STREAMING AND ANALYSIS

To get tweets via the Twitter application programming interface. The domain for this study is the Make-In-India Dataset. A Statistical Approach: This paper is divided into two parts: The two most crucial processes are: 1. Twitter data streaming, and 2. R-Studio for knowledge mining. First, the Twitter API; second, R for analyzing user sentiment. The Twitter app is required to make a connection request to the Twitter server. After a connection has been made, an authentication key is generated. Twitter's ad (data frame) file, which has been formatted as a. A usable CSV (Comma Separated Values) file is generated if the search key is "Make-In-India", and the required number of keys is provided. Opinion mining, also known as sentiment analysis1, involves collecting information from social media platforms like Twitter to conclude a topic, such as the Make-In-India program's level of support, opposition, and agnosticism among users. Here, we compare the tweet's data to a positive words file and a negative words file to determine the tweet's positive and negative scores, respectively. The final score for the tweet is found by subtracting these two numbers. the total number of tweets that have been sorted into positive, negative, and neutral sentiments to plot Make-In-popularity. India's To begin, "Make-In-India"-related tweets were collected from Twitter in the R-Studio Environment. Second, we save the extracted raw tweets in CSV format in an R database after parsing them with R based on their types. Each tweet's score is calculated and saved to a data file. In the third, we make use of the statistical program R to draw conclusions from the collected data and present our findings visually. Application/Improvements: Data on government issues, political parties, and health conditions across the country can be collected using this method. Knowing the political leaders and the program's popularity is also helpful. Sentiment analysis of tweets from customers can help you make choices.

2.5 EMOTION IN THE NEWS: A STUDY

The advent of modern technology has caused shifts in culture in a variety of spheres. Information technology (IT) advancements have allowed for faster dissemination of news and other events. It's not easy to go through all that data by hand, so experts have developed methods to do it quickly and efficiently automatically. The news covers all kinds of events, both good and bad. Sentiment analysis is a method for examining the expression of human emotion in textual data. This paper introduces a lexicon-based approach to news article sentiment analysis. Experiments performed on a BBC news dataset show the practicality and validity of the method.

Fig 2: Sample informative tweets about the disaster

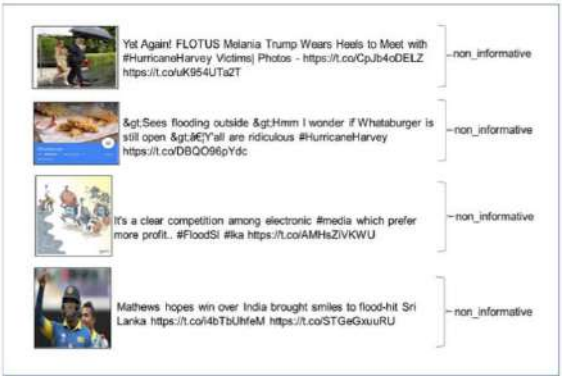

Fig 3: Sample noninformative tweets about the disaster

2. EXISTING SYSTEM

When a crisis happens, people update social media sites like Twitter immediately. Disaster relief and response teams will find this data extremely useful, enabling them to start setting priorities right away. Using text mining and machine learning techniques, we can search for words and phrases related to disasters in the vast amounts of unstructured data produced by social media platforms like Twitter. The algorithm may have trouble distinguishing between figurative and literal uses of these terms in tweets, which could lead to widespread mislabelling. This research intends to use classification models and Natural Language Processing to tell real disaster tweets from fake ones (NLP). Tweets about disasters, both real and imagined, can be found in the dataset that was obtained from the Kaggle website. Before data modeling, RStudio was used for exploratory data analysis (EDA), feature selection, and data cleaning. After that, two different testing and training schedules were compared. More so, we built classifiers with SVM, KNN, Naive Bayes, and XGBoost. When trained on the full dataset rather than sampled, SVM and XGBoost achieved accuracies of 80% and 78%, respectively. KNN (99% accuracy) and Naive Bayes (65% accuracy) both suffered from overfitting. Words like "text classification," "machine learning," and "text pre-processing" come to mind when thinking about tweets about disasters.

4. PROPOSED SYSTEM

The applicable example to categorize data with various components is the Support Vector Machine (SVM) is a helpful tool. By employing straight capability theory in high-layered space, support vector machines (SVMs) generate a (p-1) layered hyperplane in p-layered space to separate classes. SVM also makes use of the margin maximization method, which is used to select the best hyperplane for class separation. Both support vector machines (SVMs) and naive Bayes (NBayes) have proven to be highly effective and popular methods for classifying texts. Irrespective of the magnitude of the attribute area, Support Vector Machines (SVMs) exhibit a remarkable aptitude for learning.

To put it differently, if our data can be classified with a wide margin by utilizing functions from the hypothesis domain, then the ability to generalize remains achievable even when dealing with a substantial quantity of attributes. The hypothesis with the smallest VC-Dimension can be obtained by selecting the parameter that produces that hypothesis. One of the bedrocks of machine learning is the K-Nearest Neighbour (KNN) method. The algorithm's goal is to classify information according to criteria established by the machine-learned training set. KNN selects the sample groups with the highest degree of similarity based on their Euclidean distances from one another. Each term's significance in a document can be calculated using the TF-IDF technique. Text mining, natural language processing, and information retrieval are just some of the applications of this method. The weight, a statistical measure of relative importance in a document collection, is calculated using the technique.

4.1 RECORD INFORMATION AND PREPROCESS

The first step in training the SVM and KNN models is to collect the necessary training data. The data should be pre-processed to handle missing values, eliminate outliers, and normalize the data. Tweets with and without useful information are shown in Figures 2 and 3, respectively.

4.2 SEPARATE INFORMATION INTO TRAINING AND TESTING SETS.

A training set should be used to train the SVM and KNN models, while a testing set is used to assess the model's efficacy.

4.3 EXTRACTION AND SELECTION OF FEATURES

The next step is to determine what information can be gleaned from the data and what features can be extracted. Tools like principal component analysis and correlation-based feature selection are useful for this (PCA).

4.4 MODEL TRAINING USING SVM

After data pre-processing and feature extraction, the training set is ready to be used to train the SVM classifier. Gamma values, regularisation parameters, and kernel types are all adjustable in the SVM and KNN models.

4.5 MODEL EVALUATION USING THE SVM

The SVM and KNN model must be evaluated on the testing set following training. Accuracy, precision, recall, and F1-score are just a few ways the model's efficacy can be measured.

5. RESULTS

We employed Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naive Bayes, and Extreme Gradient Boosting (XGBoost) as the four supervised machine learning algorithms in our actual modeling. In the initial trial, where we split the data into 80% training and 20% testing, we utilized a subset that had imbalanced class information, with one class having more data than the other.

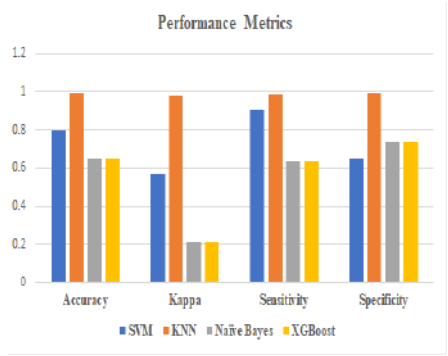

Fig 4: Performance Metrics of SVM, KNN, NB, XGBoost Algorithms

Model	Accuracy	Kappa	ensitivity	pecificity
SVM	0.7959	0.569	0.9048	0.6473
KNN	0.9906	0.9807	0.9876	0.9948
Naïve Bayes	0.6483	0.2109	0.6327	0.7379
XGBoost	0.6483	0.2109	0.6327	0.7379

Table 1 Performance metrics of SVM, KNN, NB, XGBoost

SVM and XGBoost performed well in the trial, with accuracies of 79.6% and 77.9%, respectively. KNN accuracy and Naive Bayes accuracy, on the contrary hand, were not as outstanding, both dropping below 50% (at 65% and below). We ran the simulation and three extra trials for each of the four models to guarantee an accurate representation of each group. When the proportion of participants was modified to 75/25, we re-ran the model using the original subset and uniformly dispersed sampled observations from both groups. The Table contains detailed values for each result.

7. CONCLUSION

The purpose of this project was to develop a model capable of determining whether a given text was about a tragedy. This algorithm will eventually be used by the relevant team to differentiate between genuine and fraudulent disaster reports, allowing for a more effective response. The most effective strategies for text classification were discussed in the literature review. The study showed a feasible method for solving the deterministic computational problem of spotting a potentially malicious passage of text. One of the system's biggest challenges is determining whether a given textual fragment is about a disaster. Additionally, labeled tweets were included in the operational dataset regardless of whether they referenced a real disaster. Following the acquisition of business experience, and an appreciation for the gravity of the issue at hand—we proceeded to the succeeding stages of data analysis, which included exploration, pre-processing, modeling, and evaluation. CRISP-DM was used as the method of choice (the deployment phase is anticipated). Based on the results of our four supervised machine learning models, The results indicate that the first model was overfitting, and the fourth model was not very effective (79% for KNN and 80% for SVM). Although both SVM and XGBoost performed similarly and achieved accuracy levels of 78% and 80%, respectively, we found that SVM was superior to other methods at determining which results were genuine and very accurate in distinguishing true positives and negatives.

8. REFERENCES

1. Agarwal, A., Xie, B., I. Vovsha, O. Rambow, and R.J. Passonneau Analysis of Twitter data's sentiment.
2. "Twitter Sentiment Analysis, 3-Way Classification:" by M. F. Eliktu Negative, Positive, or Neutral? Big Data (2018 IEEE International Conference on Big Data), pages 2098-2103, doi 10.1109/BigData.2018.8621970.
3. Disaster Tweets and Natural Language Processing | Kaggle 2021). Retrieved from https://www.kaggle.com/c/nlp-getting-started/data
4. on November 30, 2021. Sandhya Rani, K., and Vasudha Rani, V. Indian Journal of Science

and Technology, 9(45), Twitter Streaming and Analysis Using R. doi:
5. 10.17485/ijst/2016/v9i45/97914 Antony Samuels and John McGonigal 2020). Analysis of News Sentiment.
6. Mathias S. Detecting fake news from twitter, Feb. 2018.
7. Dabreo S. Comparative analysis of fake news detection using machine learning and deep learning techniques, Apr. 2020
8. Vaishnavi R. Fake news detection/fake buster, May 2020.
9. Fernandez N. A deep neural network for fake news detection, Nov. 2017.
10. Ranjan E. Fake news detection by learning convolution filters through contextualized attention, Aug. 2019.
11. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. Fakenewsnet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data. 2020;8(3):1–5.
12. Rashed Ibn Nawab M et al. Rumour detection in social media with user information protection. EJECE 2020;4(4).
13. Krishnan S, Chen M. Identifying tweets with fake information. In: Proceedings of the international conference on information reuse and integration; 2018 Jul 6–9. Salt Lake City, USA: IEEE; 2018. p. 460–4.
14. Aphiwongsophon S, Chongstitvatana P. Identifying misinformation on Twitter with a support vector machine. EASR, Mar. 2020 Sheryl Mathias, detecting fake news from Twitter, Feb. 2018.
15. Nyow NX, Chua HN. Detecting fake news with tweets' properties. IEEE AINS., Nov. 2019.