A Review on Text Sentiment Analysis using Disaster Allied Tweets

Selvakumar G1, Adithi Mookambal2, Madhumidha3

School of Computer Science and Engineering (SCOPE)1,2,3

Vellore Institute of Technology1,2,3

Chennai, India

# **Abstract**

Social media platforms have played a key role in the expansion of information exchange and as a significant source of news for all types of industries and personal matters over the last decade. We can better understand the current opinions and values that people have around them by extracting and analyzing data from social media, and we can help the government or any other organization that analyses this type of data and approaches a solution to any problems that occur to keep an ideal around us. Modern algorithms from machine learning, deep learning, and transfer learning technologies can be used to analyze and learn more about data as it becomes more widely available, resulting in better outcomes.

***Keywords: Tweets, Sentiment Analysis***

# **Introduction**

Twitter is a social media platform that allows users to send and receive short messages known as tweets. Tweets can be up to 280 characters long and can include text, images, or links. Tweets can also be interacted with by liking, commenting, or retweeting them. There is no specific data available on the average number of tweets posted on Twitter each day because this varies depending on a few factors such as the number of users, the events and trends being discussed, and the time of year. Twitter, on the other hand, is estimated to receive millions of tweets per day from its users. With over 330 million monthly active users as of the fourth quarter of 2021, Twitter has a large and active user base. The platform is especially popular for news and information sharing, as well as discussions about current events, politics, and pop culture. Businesses and organizations also frequently use it for marketing and customer service.

Twitter data has been widely used for various purposes in artificial intelligence (AI) and machine learning (ML). Among the most common applications are

1. Sentiment analysis: Machine learning models have been trained using Twitter data to classify the sentiment expressed in tweets (positive, negative, or neutral). This has a wide range of applications, including marketing, customer service, and political analysis.
2. NLP: Twitter data has been applied to train machine learning models for tasks like language translation, text summarization, and text classification. Twitter data was utilized to investigate patterns and trends in social interactions, such as how information spreads on the platform and how users connect with one another.
3. Trend analysis: Twitter content is used to assess trends and patterns in popular topics and hashtags on the platform.
4. Event detection: Twitter data is being used in real-time to identify and track events and trends such as natural disasters and major news events.

Overall, Twitter's large volume of data and real-time nature make it a valuable resource for AI and ML applications, especially in natural language processing and social media analysis.

Twitter data has frequently been used to study and comprehend the social impact of pandemics and natural disasters. For example, researchers have examined how people cope with and respond to the COVID-19 pandemic, including how they share information and express their emotions, using Twitter data. Similarly, Twitter data has been applied to investigate how people react to natural disasters such as hurricanes and earthquakes, as well as how they seek and aid one another.

There are several ways that Twitter data can be clouded or biased during pandemics and disasters. One issue is that not everyone has equal access to the internet or social media, so the data may not be representative of the entire population. Additionally, people may be more likely to tweet about certain topics during a pandemic or disaster, which can skew the data. For example, during the COVID-19 pandemic, there may be more tweets about the virus and its impact on daily life, which could make it difficult to accurately analyze trends and patterns.

During pandemics and disasters, Twitter data can be clouded or biased in a variety of ways. The data may not be representative of the entire population because not everyone has equal access to the internet or social media. Furthermore, during a pandemic or disaster, people may be more likely to tweet about specific topics, which can skew the data. During the COVID-19 pandemic, for example, there may be more tweets about the virus and its impact on daily life, making it difficult to accurately analyze trends and patterns.

# **Literature Survey**

Silva et al. present a method for analyzing tweet sentiment with classifier ensembles. Classifier ensembles are machine learning models that combine multiple individual classifier predictions to make more accurate predictions. They began by gathering a dataset of tweets annotated with their sentiment (positive, negative, or neutral). The tweets were then preprocessed by removing punctuation and stop words and converting them to lowercase. The authors then used the dataset to train a variety of individual classifiers, including support vector machines (SVMs), naive Bayes classifiers, and decision tree classifiers. They also trained several classifier ensembles using various methods for combining individual classifier predictions, such as majority voting and weighting. They then used a variety of evaluation metrics to assess the performance of individual classifiers and classifier ensembles on the dataset. These metrics included accuracy, precision, and recall. They discovered that classifier ensembles outperformed individual classifiers, especially when weighting was used to combine the predictions. It was also mentioned that classifier ensembles can be an effective method for tweet sentiment analysis and can improve sentiment prediction accuracy when compared to individual classifiers [1].

Hwang et al. describe a method for performing aspect-level sentiment analysis, which entails analyzing sentiment expressed toward specific aspects or features of a product or service. Businesses can use this to better understand customer feedback and improve their products. To perform aspect-level sentiment analysis, the authors propose a model that combines a convolutional neural network (CNN) with a BERT-GCN (Bidirectional Encoder Representations from Transformers with Graph Convolutional Networks). BERT is a language model capable of understanding the context and meaning of words in a sentence, whereas GCN is a type of neural network capable of learning from graph-structured data. They began by gathering a dataset of customer reviews that had been annotated with the aspects mentioned in each review as well as the sentiment expressed toward each aspect. They then preprocessed the data by tokenizing the reviews and using BERT to generate word embeddings. The CNN-BERT-GCN model was then trained on the dataset by the authors. The model generates word embeddings for each review using BERT, and the GCN module then processes the word embeddings to identify the aspects mentioned in the review. The CNN module is then used to categorize the sentiment expressed in response to each identified aspect. The authors used a variety of evaluation metrics to assess the model's performance on the dataset, including accuracy, precision, and recall. They discovered that the model performed well on the task of aspect-level sentiment analysis. Finally, the authors demonstrated that the CNN-BERT-GCN model can perform aspect-level sentiment analysis effectively and can be a useful tool for businesses to understand customer opinions and improve their products [2].

Onyenwe et al. present a study that uses data from the 2017 Anambra State Governor election in Nigeria to investigate the relationship between sentiment expressed in tweets and election results. They gathered a dataset of election-related tweets that were annotated with the sentiment expressed in each tweet (positive, negative, or neutral). They then ran sentiment analysis on the tweets to determine how people felt about each political party and candidate. The authors discovered a positive correlation between the sentiment expressed in tweets and election results, implying that parties and candidates who received more positive sentiment in tweets fared better in the election. They also discovered that tweet sentiment was a better predictor of election results than traditional indicators like campaign spending and voter turnout. The authors also conducted a qualitative analysis of the tweets, which revealed that campaign promises, party reputation, and personal characteristics of the candidates all influenced the sentiment expressed in tweets. They also demonstrated that sentiment analysis of tweets can be a useful tool for predicting election results, and that tweet sentiment is influenced by a variety of factors. This data could help political parties and candidates understand how to appeal to voters and increase their chances of winning elections [3].

SentiCircles is a model proposed by Saif et al that consists of three components: a word embedding module, a context module, and a concept module. The word embedding module converts tweet words into numerical representations known as word embeddings, which capture the meaning of the words. The context module considers the context in which the words appear, whereas the concept module considers the overall meaning or concept expressed in the tweet. The authors gathered a dataset of tweets annotated with their sentiment to train the SentiCircles model (positive, negative, or neutral). The tweets were then preprocessed by removing stop words and converting them to lowercase, and word embeddings were generated using a pre-trained language model. The authors assessed the SentiCircles model's performance on the dataset using a variety of evaluation metrics, including accuracy, precision, and recall. They discovered that the model outperformed other innovative sentiment analysis models, particularly in correctly identifying neutral sentiment. The authors demonstrated that the SentiCircles model is a useful tool for performing semantic sentiment analysis on tweets and can improve sentiment prediction accuracy when compared to other approaches. Because it can consider the context and meaning of words in tweets, it is especially well-suited to understanding the sentiment expressed in social media messages [4].

Yao et al. present a study that looks at the sentiment expressed in tweets from different megacities (cities with more than 10 million people) during the COVID-19 pandemic. The paper's authors compiled a dataset of tweets from ten global megacities, including New York, London, and Tokyo. The tweets were then preprocessed by removing stop words and converting them to lowercase, and word embeddings were generated using a pre-trained language model. They then used the dataset to train a variety of machine learning models, including support vector machines (SVMs), naive Bayes classifiers, and decision tree classifiers. They also trained several classifier ensembles using various methods for combining individual classifier predictions, such as majority voting and weighting. The authors then used a variety of evaluation metrics to assess the models' performance on the dataset, including accuracy, precision, and recall. They discovered that classifier ensembles outperformed individual classifiers, and that the models achieved consistent performance on the sentiment analysis task. They also conducted a qualitative analysis of the tweets to determine the most common topics and themes in each megacity. They discovered that the most common themes were related to the COVID-19 pandemic, such as information about the virus, control measures, and the impact on daily life. Finally, they demonstrated that machine learning techniques can be used to perform sentiment analysis on tweets, and that sentiment expressed in tweets from various megacities during the COVID-19 pandemic was influenced by the pandemic itself and its impact on daily life. This data could be useful in understanding how different populations around the world are dealing with the pandemic and the measures put in place to combat it [5].

Chakraborty et al. proposed a method to demonstrate how popularity affects accuracy in social media presents a study that investigates the performance of deep learning classifiers on the task of sentiment analysis of COVID-19 pandemic tweets. The paper's authors also investigate how the popularity of assorted topics affects the accuracy of the classifiers. They gathered the data by searching for tweets on the social media platform Twitter with a set of keywords related to the COVID-19 pandemic. The sentiment of the tweets was then annotated (positive, negative, or neutral). They also preprocessed the tweets by removing stop words and converting them to lowercase and used a pre-trained language model to generate word embeddings. They then used the dataset to train a variety of deep learning classifiers, including long short-term memory (LSTM) networks and convolutional neural networks (CNNs). The classifiers' performance on the dataset was then evaluated using a variety of evaluation metrics, including accuracy, precision, and recall. They discovered that LSTM networks outperformed CNNs in general, and that classifiers performed well in sentiment analysis. The authors also investigated how the popularity of assorted topics in tweets affected the classifiers' accuracy. They discovered that the classifiers performed better when the tweets were about more popular topics and performed worse when the tweets were about less popular topics. Finally, they demonstrated that deep learning classifiers can be effective for performing sentiment analysis on tweets about the COVID-19 pandemic, and that the popularity of different topics can affect the classifier's accuracy [6].

**Analysis & Findings**

Twitter is a major data source for determining current public sentiment and analyzing it using innovative techniques. However, gathering data through web scraping is difficult, and the variety of topics in tweets will affect the study's generalization. So, target what type of data is required and categorize recent tweets based on it before passing them through the model. This data must be vectorized and embedded, as well as the necessary data preprocessing steps. The most used model for this type of sentiment analysis is recurrent neural network models such as GCN and LSTM, which provide better classification metrics and can be taken to the edge by pushing this collected resource into a Transfer learning model like Universal Sentence Encoder and test its performance.

**Conclusion**

The process of automatically analyzing text data to determine the sentiment or emotion expressed in it is known as text sentiment analysis. Deep learning techniques, such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, have grown in popularity for text sentiment analysis because of their ability to learn complex patterns and relationships in data. Pay close attention to context and meaning: Researchers have concentrated on developing models that consider the context and meaning of words in a text, rather than just the words themselves. This is referred to as semantic sentiment analysis. And, because it can speed up training and improve performance, the use of transfer learning, which involves using pre-trained models on large datasets as a starting point for training on a new task, has become a popular approach in text sentiment analysis. Overall, these trends reflect a focus on improving the accuracy and robustness of text sentiment analysis models and developing models capable of better understanding the text's context and meaning.

# **References**

1. Nádia F.F. da Silva, Eduardo R. Hruschka, Estevam R. Hruschka, "Tweet sentiment analysis with classifier ensembles", Decision Support Systems, Volume 66, 2014, Pages 170-179.
2. H. T. Phan, N. T. Nguyen, and D. Hwang, "Aspect-Level Sentiment Analysis Using CNN Over BERT-GCN," in IEEE Access, vol. 10, pp. 110402-110409, 2022.
3. Onyenwe, I., Nwagbo, S., Mbeledogu, N. et al. The impact of political party/candidate on the election results from a sentiment analysis perspective using #AnambraDecides2017 tweets. Soc. Netw. Anal. Min. 10, 55 (2020).
4. Saif, H., Fernandez, M., He, Y., Alani, H. (2014). SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter. In: Presutti, V., d’Amato, C., Gandon, F., d’Aquin, M., Staab, S., Tordai, A. (eds) The Semantic Web: Trends and Challenges. ESWC 2014. Lecture Notes in Computer Science, vol 8465. Springer, Cham.
5. Zhirui Yao, Junyan Yang, Jialin Liu, Michael Keith, ChengHe Guan, "Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19", Cities, Volume 116,2021,103273.
6. Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, Aboul Ella Hassanien, “Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media”, Applied Soft Computing, Volume 97, Part A, 2020, 106754.