Crisis Detection Through Twitter

Adya Agarwal (2022043)
Mehul Pahuja (2022295)
Rahul Jha (2022389)
Group number 22

ML-EndSem Project Presentation

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

- The is a need for timely information during crisis situations to ensure effective response and situational awareness.
- By utilizing machine learning to automatically classify crisis-related tweets, this project aims to enhance disaster management strategies, enabling faster and more accurate detection of crises and ultimately supporting real-time decision-making during emergencies.
- This has several applications in crisis, journalism, research, and many more.

Literature review

1. Ashktorab et al. (2014) - Tweedr

- **Method**: Logistic Regression, Random Forests & Gradient Boosted Trees for disaster tweet classification.
- **Key Finding**: Improved precision/recall with domain-specific features (n-grams, POS tags, keywords).

Method	F1	Pr	Re	Acc	AUC
LogReg	**.65 ± .07**	.78 ± .08	.57 ± .09	**.86 ± .03**	**.88**
NB	.63 ± .06	.55 ± .07	**.75 ± .09**	.80 ± .03	.84
DTree	.54 ± .09	**.93 ± .07**	.39 ± .09	.85 ± .02	.69
KNN	.51 ± .04	.83 ± .10	.38 ± .04	.84 ± .02	.73
sLDA	.50 ± .07	.42 ± .06	.65 ± .15	.70 ± .05	.77

2. Chaudhari & Govilkar (2015)

- **Method**: Overview of sentiment analysis techniques for social media text.
- **Key Finding**: Emphasized challenges in feature selection and classification.

3. Nguyen et al. (2016)

- **Method**: CNN-based model for disaster tweet classification.
- **Key Finding**: Achieved high F1 score, proving CNN's effectiveness for short, informal texts.

Dataset description

The dataset comprises a total of 247,000 tweets, categorized into crisis and non-crisis classes. It was created by combining multiple publicly available sources, including:

1. Twitter Data on Disaster-Related Tweets (228,005 tweets): Sourced from the Omdena platform, which provides data for social good projects.
2. Stepanenko Disaster Tweets Dataset (11,380 tweets): Available on Kaggle, focused on tweets related to disaster events.
3. Kaggle NLP Getting Started Competition Dataset (8,562 tweets): Initially used for a natural language processing competition focused on disaster-related tweets.

Dataset description

- We visualized the distribution of crisis and non-crisis tweets in the dataset.
- We explored the distribution of tweet lengths, as shown in Figure 2.

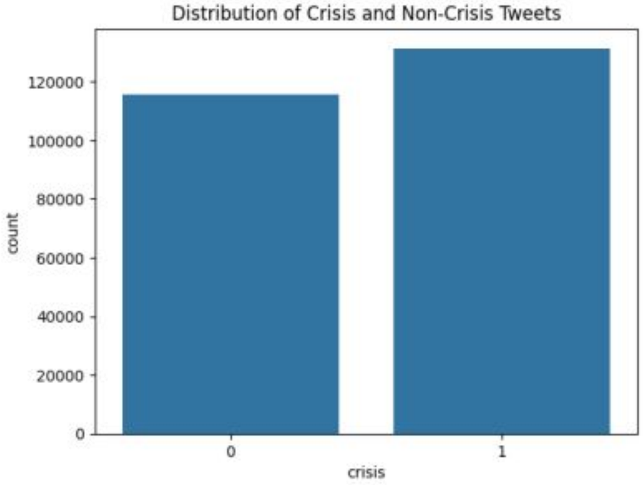


Figure 1: Distribution of Crisis and Non-Crisis Tweets

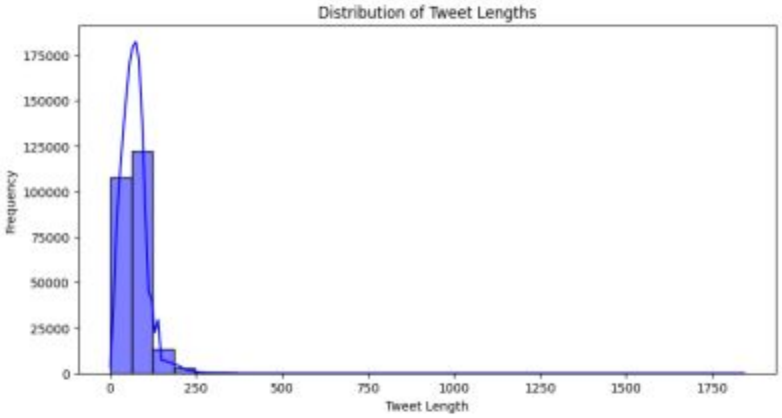


Figure 2: Distribution of Tweet Lengths

Dataset description

- To better understand the sentiment in tweets, we plotted the distribution of sentiment polarity scores in Figure 4.
- We also compared sentiment polarity between crisis and non-crisis tweets.

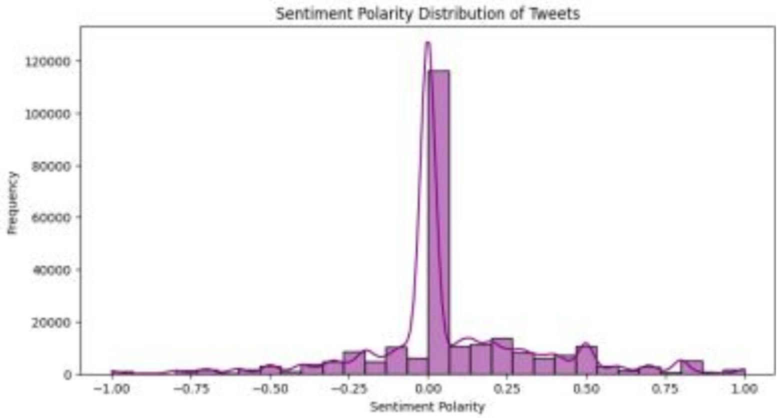


Figure 4: Sentiment Polarity Distribution of Tweets

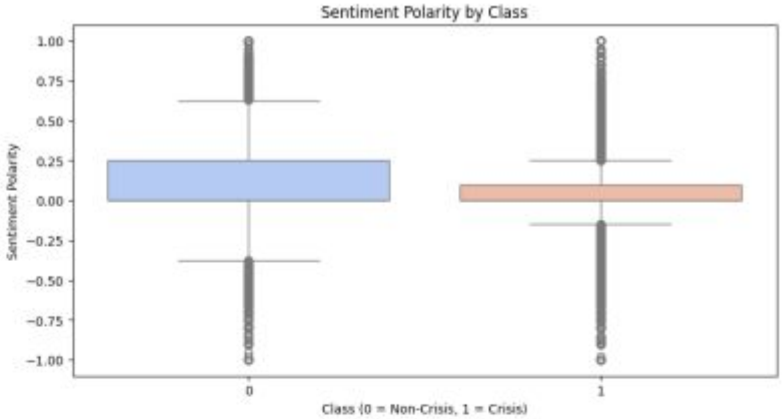


Figure 5: Sentiment Polarity by Class

Dataset description

- We analyzed the top 10 most common bigrams (two word phrases) for crisis and non-crisis tweets.

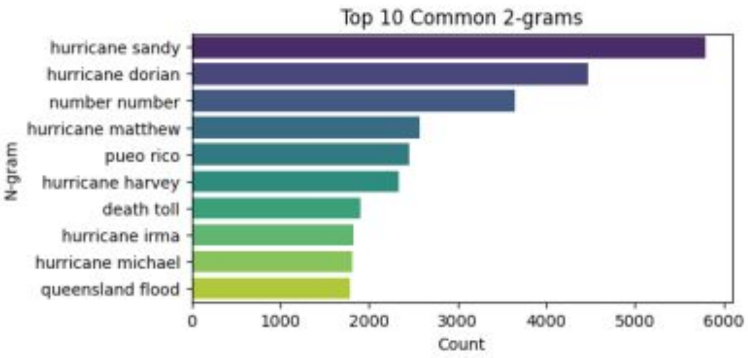


Figure 6: Top 10 Common Bigrams for Crisis Tweets

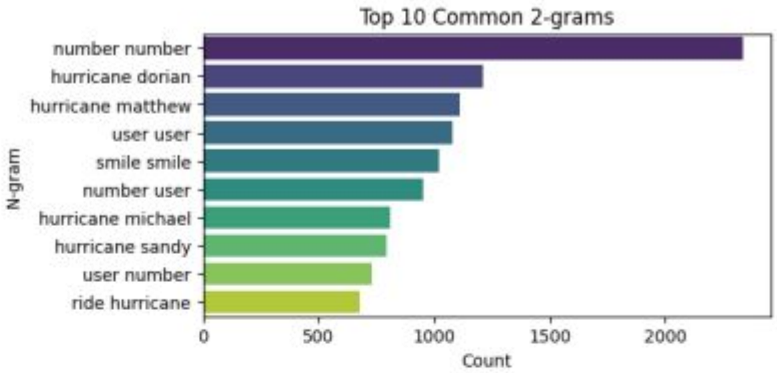


Figure 7: Top 10 Common Bigrams for Non-Crisis Tweets

Methodology

Machine Learning Models:

- **Logistic Regression**: A basic linear model to classify tweets based on extracted features.
- **Decision Tree**: A tree-based model used to capture non-linear relationships between features.
- **Random Forest**: An ensemble method that improves classification performance by averaging multiple decision trees.
- **XGBoost**: An optimized gradient boosting algorithm designed for efficient and accurate classification.
- **AdaBoost**: A boosting technique that focuses on hard-to-classify samples to improve model performance.
- **K-Nearest Neighbors (KNN)**: A simple, instance-based learning algorithm for classifying tweets based on proximity in feature space.
- **Support Vector Machine (SVM)**: A robust classifier capable of handling high-dimensional feature spaces, particularly useful for text classification.
- **Multilayer Perceptron (MLP)**: A feedforward neural network used to capture complex patterns in the data.
- **Convolutional Neural Network (CNN)**: A deep learning model designed to extract spatial and local features from tweet text using convolutional layers.

Methodology

Feature Extraction and Processing:

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: Used to transform text data into numerical form, highlighting the importance of words in context.
- **PCA (Principal Component Analysis)**: Dimensionality reduction technique to enhance model efficiency by reducing the feature space.

Model Tuning:

- **Grid Search**: A technique to fine-tune the hyperparameters of each model for optimal performance.

Exploratory Data Analysis (EDA):

- **Tweet Length Distribution**: Visualizing the distribution of tweet lengths to understand the nature of the data.
- **Sentiment Polarity**: Analyzing the distribution of sentiment polarity to examine the emotional tone of tweets.
- **Common Bigrams**: Identifying the most frequent two-word phrases in crisis and non-crisis tweets to uncover meaningful patterns.
- **Sentiment Polarity by Class**: Comparing sentiment polarity between crisis and non-crisis tweets to discern patterns in emotional content.

Results/Analysis/conclusion

Traditional Machine Learning Models

- **Logistic Regression**: F1 score of **91.07%**, serving as a strong baseline.
- **Decision Tree**: F1 score of **90.87%**, effectively capturing non-linear patterns.
- **Random Forest**: **92.24%** F1 score, demonstrating robustness and handling complexity well.
- **XGBoost**: F1 score of **91.11%**, excelling with large feature spaces and high efficiency.
- **Support Vector Machine (SVM)**: Outperformed other models with an F1 score of **91.43%**.
- **K-Nearest Neighbors (KNN)**: Lower F1 of **78.83%**, but effective in specific settings, improving to **84.21%** with PCA.

Model	Embedding	F1 score (Test)
Logistic Regression	TF-IDF	0.9107
Logistic Regression	TF-IDF + PCA	0.9015
Decision Tree	TF-IDF	0.9087
Decision Tree	TF-IDF + PCA	0.8233
Random Forest	TF-IDF	0.9224
Random Forest	TF-IDF + PCA	0.8909
XGBoost	TF-IDF	0.9111
XGBoost	TF-IDF + PCA	0.9036
AdaBoost	TF-IDF	0.8876
AdaBoost	TF-IDF + PCA	0.8536
K-Nearest Neighbors (KNN)	TF-IDF	0.7883
K-Nearest Neighbors (KNN)	TF-IDF + PCA	0.8421
Support Vector Machine (SVM)	TF-IDF	0.9143
Support Vector Machine (SVM)	TF-IDF + PCA	0.8099
MLP	TF-IDF	0.9268
MLP	TF-IDF + PCA	0.9200
CNN	TF-IDF	0.8968
CNN	TF-IDF + PCA	0.8900

Table 1. Comparison of Model Accuracies with and without PCA

Results/Analysis/conclusion

Deep Learning Models:

- **CNN**: Achieved **89.2%** F1 score, strong but weaker than ensemble models.
- **MLP**: Accuracy of **92.68%**, showing strong classification performance, though slightly behind ensemble methods.

Baseline Models:

- **Random Forest** (without TF-IDF): **65.91%** accuracy using basic features like tweet length and sentiment.
- **KNN** with TF-IDF: F1 score of **78.83%**, improved to **84.21%** with PCA.

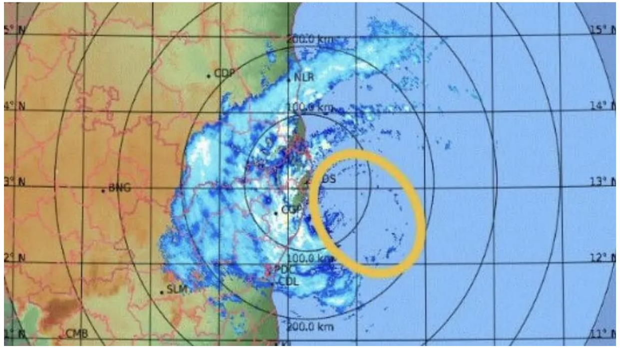
Model	Embedding	F1 score (Test)
Logistic Regression	TF-IDF	0.9107
Logistic Regression	TF-IDF + PCA	0.9015
Decision Tree	TF-IDF	0.9087
Decision Tree	TF-IDF + PCA	0.8233
Random Forest	TF-IDF	0.9224
Random Forest	TF-IDF + PCA	0.8909
XGBoost	TF-IDF	0.9111
XGBoost	TF-IDF + PCA	0.9036
AdaBoost	TF-IDF	0.8876
AdaBoost	TF-IDF + PCA	0.8536
K-Nearest Neighbors (KNN)	TF-IDF	0.7883
K-Nearest Neighbors (KNN)	TF-IDF + PCA	0.8421
Support Vector Machine (SVM)	TF-IDF	0.9143
Support Vector Machine (SVM)	TF-IDF + PCA	0.8099
MLP	TF-IDF	0.9268
MLP	TF-IDF + PCA	0.9200
CNN	TF-IDF	0.8968
CNN	TF-IDF + PCA	0.8900

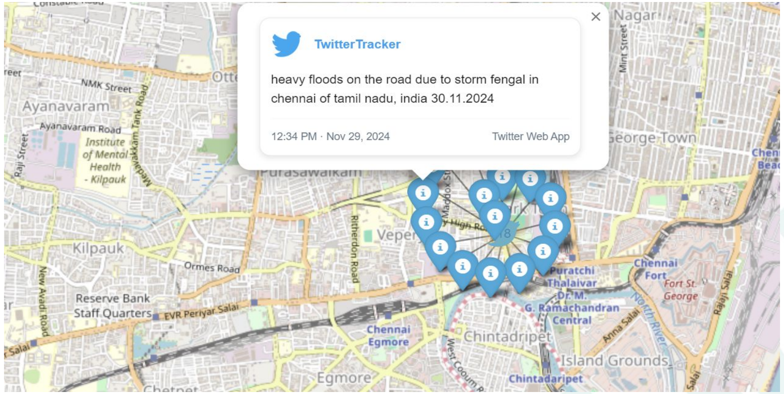
Table 1. Comparison of Model Accuracies with and without PCA

Interface

Crisis Monitoring: Cyclone Alert in Chennai

- **Cyclone Update**:
  A cyclone is approaching Chennai, bringing heavy rain, strong winds, and potential flooding. Timely updates are crucial for effective disaster response.
- **Real-Time Crisis Detection**:
  Our ML-powered tool will classify live tweets about the cyclone, flagging crisis-related information such as evacuations and safety updates.
- **Impact**:
  The interface will provide real-time insights, helping decision-makers stay informed and respond swiftly to the evolving situation.





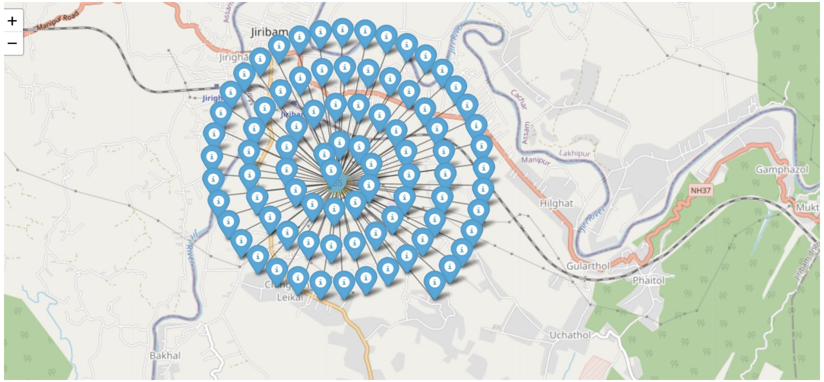
Interface

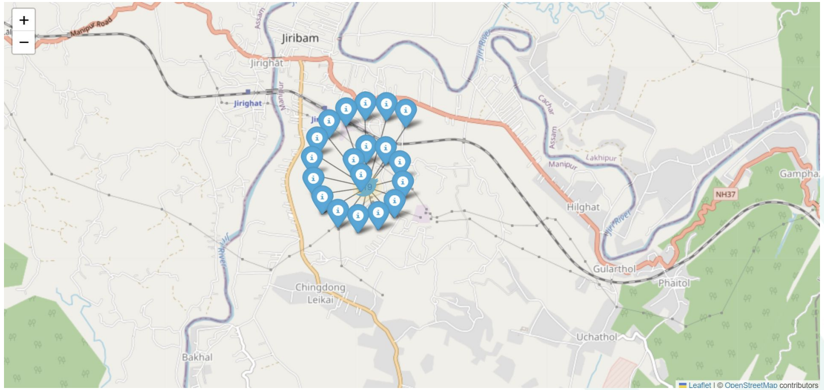
The disaster classified tweets collected from the IP addresses are marked on a map, and can be seen by the user.

The above image represents all tweets from the following latitude and longitude.

24.6637° N, 93.9063° E

The below image represents the crisis tweets after running the classification model.





Thank You