Disaster-Related Tweet Classification Using NLP

I. Introduction

A. Background

Social media platforms, particularly Twitter, have become critical sources of information during disaster events. Individuals often share updates about ongoing situations, post calls for help, and offer insights into real-time conditions. These platforms provide a wealth of data, but the sheer volume of tweets makes manual filtering and analysis impractical.

Automated systems leveraging natural language processing (NLP) can help by identifying disaster-related tweets, enabling timely and informed decision-making for disaster response and management.

B. Problem

Many tweets use metaphorical phrases that involve terms that are related to disasters, but they are nothing more than a figure of speech. This in certain cases can be misleading. The main challenge addressed in this project is to accurately classify tweets as disaster-related or non-disaster-related.

Tweets present several unique difficulties, such as their informal language, use of abbreviations, and variability in structure. Traditional machine learning models often fall short in understanding the nuanced relationships between words and their contexts, resulting in suboptimal classification performance. An efficient classification model must address these issues while maintaining high accuracy to be practically useful in disaster scenarios.

C. Importance

A system capable of identifying disaster-related tweets has far-reaching implications for disaster response teams, governments, and humanitarian organizations. Such a system can:

- **Enable resource prioritization:** Quickly identify areas in need of urgent attention.
- **Support real-time decision-making:** Provide actionable insights from the vast sea of social media data.
- **Facilitate disaster planning and recovery:** Use classified data to improve response strategies for future events. By automating this process, these systems can help save lives, reduce disaster impacts, and enhance coordination among stakeholders.

D. Existing Literature

Research in text classification has evolved significantly over the years, transitioning from simpler methods like TF-IDF to sophisticated transformer models such as BERT. TF-IDF (Term Frequency-Inverse Document Frequency) has been a popular choice due to its simplicity and effectiveness in capturing the importance of terms
within a document. However, it lacks the ability to capture semantic relationships between words. Continuous Bag of Words (CBOW), a neural network-based embedding model, has addressed this limitation by learning word representations based on their surrounding context. Nevertheless, CBOW struggles with capturing complex word dependencies.

BERT (Bidirectional Encoder Representations from Transformers) represents a significant advancement in NLP. Its bidirectional approach allows it to understand the context of a word in relation to both preceding and succeeding words. While studies have demonstrated its effectiveness across various NLP tasks, its application in disaster-related tweet classification is a relatively unexplored area that promises to yield improved results.

E. System Overview

This project proposes a system to classify tweets into two categories: disaster-related and non-disaster-related. The system

consists of a robust data mining pipeline, which follows these key steps:

1. **Data Preprocessing:** Tweets are cleaned (handling missing values), tokenized, and standardized to remove noise and inconsistencies.

2. **Feature Extraction:** Different embedding models (TF-IDF, CBOW, and BERT) are applied to convert textual data into numerical formats suitable for machine learning models.

3. **Model Training:** Classifiers such as Logistic Regression, Support Vector Machines (SVM), and Decision Trees are trained on these embeddings to identify the best-performing model.

4. **Evaluation:** The models are assessed using metrics such as Precision, Recall, and F1-score to ensure accuracy and reliability.

F. Data Collection

The primary dataset for this study is sourced from Kaggle's "NLP Getting Started" competition, which contains labeled tweets categorized as disaster-related or non-disaster-related. This dataset serves as an excellent foundation for the project, offering a balanced representation of relevant and irrelevant tweets.

To ensure that the dataset is original and not biased or skewed, more data points were added by each team member. By leveraging this dataset, the system aims to simulate real-world scenarios where distinguishing critical information is essential.

G. Components of the ML System

After collecting the data and cleaning it, the next stage was to pass it into a pipeline that would generate the required results. The machine learning system incorporates the following components:

1. **Feature Extraction:**

Feature extraction is used to get the embeddings for each word present in the tweets. This is a necessary step to allow the model to understand the nuances between the words and classify accurately. The following text embedding models were used:

1. **TF-IDF:** A classic text representation technique that quantifies the importance of terms in a document. It forms a sparse matrix representation of text data and has been tested with models like Logistic Regression and KNN in this project.

2. **CBOW:** A word embedding model that predicts a target word based on its surrounding context. In this project, 100-dimensional embeddings were generated, with a context window size of 5 and 300 training epochs.

3. **BERT:** A transformer-based embedding model that captures complex relationships between words by considering both preceding and succeeding contexts. It outputs dense embeddings, which are fed into machine learning classifiers for superior performance.

2. **Model Training:**

After getting the embeddings for the entire dataset, they were trained and tested on a series of models (Logistic Regression, Decision Trees, Random Forest, naïve Bayes, and SVMs).

The results from each of the models, namely the re-call precision and accuracy were used for the comparative analysis.

H. Experimental Results

Initial experiments revealed notable differences in the performance of these embedding methods:

• **TF-IDF:** Logistic Regression performed the best, achieving an F1-score of approximately 75%, while KNN delivered the least accurate results. The overall average F1-score was around 70%, indicating moderate performance.

• **CBOW:** Although t-SNE visualizations demonstrated effective clustering of disaster-related and non-disaster related tweets, the precision and recall scores were lower, indicating room for improvement.

• **BERT:** The use of BERT embeddings significantly enhanced classification performance, with SVM achieving the highest F1-score (~78%). This aligns with the hypothesis that advanced embeddings capture intricate word relationships better than traditional models.

The results suggest that transformer-based models like BERT are highly effective for disaster-related tweet classification,

warranting further exploration and fine-tuning.

II. Important definitions

The project focuses on analyzing social media communications through tweets during emergencies, to identify real disaster situations for better resource allocation during disasters and faster emergency response. Following are the important definitions:

A. Data Description

A comprehensive dataset containing 7,613 unique social media posts from Twitter (X) forms the foundation of this analysis. Each tweet contains text content that may or may not indicate a real disaster. There are approximately 4,000 non-disaster related posts and 3,000 disaster-related tweets.

B. Prediction Target

The classification system employs a binary target variable where:

- A value of 1 indicates genuine disaster-related content
- A value of 0 represents non-disaster content

C. Key Variables

Following are the important elements present in the dataset:

- Tweet text: The primary content requiring analysis
- Processed text: Normalized version after removing special characters and converting to lowercase - Keywords: Essential terms extracted through TF-IDF methodology

D. Problem Statement

Context and Given Information

Emergency response organizations and news agencies increasingly rely on social media for real-time disaster monitoring. However, the challenge lies in distinguishing between literal and metaphorical usage of disaster-related terms in social media communications.

Project Objectives

The primary aim is to develop a system that can:

- Accurately classify disaster-related communications
- Implement robust natural language processing techniques
- Create a comprehensive data mining pipeline for reliable predictions

Technical Constraints

The project faces several technical hurdles:

- Text processing challenges due to informal language patterns and social media-specific content - Limited context availability due to the brief nature of tweets
- Complex disambiguation requirements for metaphorical language

Modeling Considerations

The implementation must address:

- Careful balance between false positive and false negative classifications
- Management of class imbalance in the dataset
- Effective processing of informal text while maintaining contextual understanding

This framework provides a foundation for developing an efficient disaster management system through text mining and classification techniques.

III. Overview of Proposed Approach/System

The pipeline from data collection to model evaluation is the main focus of this part, which also describes the methodical approach we took for our project. Every step has been planned to guarantee thorough study and performance comparison of different NLP models.

A. Information Gathering

The initial phase is combining information from many sources, including our own carefully selected input tweets and publicly accessible Kaggle datasets. The dataset is guaranteed to be complete and indicative of actual text situations thanks to its multi-source data aggregation.

B. Data Purification

A crucial step in ensuring high-quality input for analysis is data pretreatment. It consists of eliminating duplicate entries and dealing with missing values.

Text can be cleaned by removing extraneous letters, punctuation, and URLs.

C. Feature Extraction

To represent textual data numerically, we explore different embedding techniques:

- TF-IDF: A traditional approach that emphasizes term frequency while down-weighting common terms. BERT and

RoBERTa: Advanced transformer-based models that capture deep semantic relationships and contextual meanings in the text. Tokenized sentences from the dataset are used to train word embeddings with a 100-dimensional representation. Words appearing at least once are considered, with a context window size of 5. The Continuous Bag of Words (CBOW) architecture is used during training for 300 epochs.

D. Model Building

Multiple machine learning models are tested to identify the best performing architecture for our use case:

- Logistic Regression
- Decision Trees
- Naive Bayes Classifier
- Random Forest
- Support Vector Machines (SVMs)

These models are evaluated on their ability to generalize well on the dataset, focusing on predictive accuracy and interpretability.

IV. Technical Details of Proposed Approaches/System

A. Data Collection

The dataset used for this project combines a primary source from Kaggle's *"NLP Getting Started"* competition with additional tweets contributed by team members. The data collection process involved the following steps:

1. **Primary Dataset:**

o The Kaggle dataset provided labeled tweets, categorizing them as disaster-related or non-disaster-related. o It served as the foundation for training and evaluating the models due to its balanced representation and established labeling.

2. **Supplementary Data:**

o To enhance the diversity of the dataset and maintain originality, each team member curated and labeled their own tweets.

o These additional tweets reflect various linguistic styles, local contexts, and emerging disaster scenarios, ensuring the dataset is representative of real-world variability.

o The expanded dataset contributes to the model's ability to generalize across different disaster tweet types.

The resulting dataset offers a balanced and comprehensive set of tweets, ensuring robust training and evaluation for the classification models.

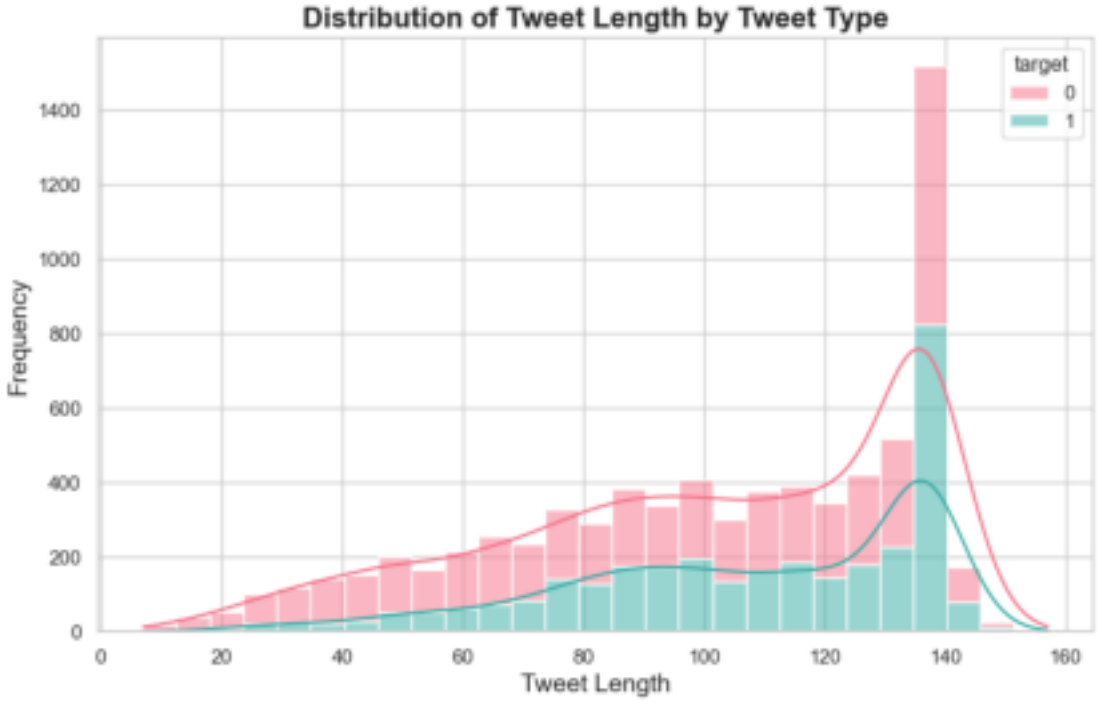

Figure 1.
Distribution of Tweet Length by Tweet Type

B. Feature Extraction

Multiple approaches were used to get the features from the processed dataset. These features are word embeddings that allow the model to understand the relationship between the various words in human language and train it to accurately classify the target word. The project incorporates three approaches:

TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF assigns importance to words based on their frequency in a tweet relative to the entire dataset. • **Process:**

o Calculates the term frequency (TF) and inverse document frequency (IDF) for each word. o Combines these scores to create a sparse matrix representing the importance of terms in tweets.

• **Implementation:**

o Applied to the entire dataset to generate features.

o Simple and computationally efficient, suitable for initial experiments.

• **Limitations:**

o Unable to capture word semantics or contextual relationships, limiting its performance in nuanced text classification tasks.

CBOW (Continuous Bag of Words)

CBOW is a neural network-based approach that generates dense vector representations for words. • **Process:**

o Predicts a target word based on its surrounding context words within a fixed window size.

o Converts words into dense, 100-dimensional embeddings that represent their semantic and contextual meaning.

- **Implementation:**
  - o Context window size set to 5, trained over 300 epochs.
  - o Only words occurring at least once are included to reduce noise.
- **Strengths and Limitations:**
  - o Effective for capturing local word relationships but struggles with global context and complex dependencies.

BERT (Bidirectional Encoder Representations from Transformers)

BERT is a state-of-the-art model that captures deep contextual relationships in text using a bidirectional transformer architecture.

- **Process:**
  - o Converts tweets into tokenized sequences with added [CLS] and [SEP] tokens.
  - o Outputs dense embeddings for each word, with the [CLS] token embedding used as the feature vector for tweet classification.
- **Implementation:**
  - o Pre-trained BERT model applied to generate embeddings for all tweets.
  - o Focuses on the entire context, resulting in high-quality feature representations.
- **Advantages:**
  - o Superior semantic understanding and context sensitivity, leading to improved model accuracy.

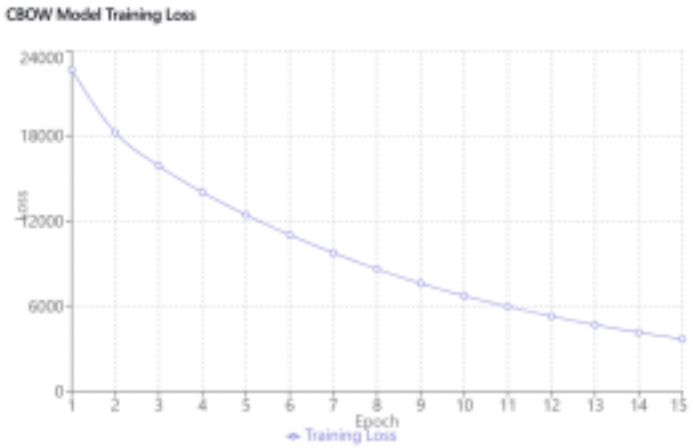

Figure 2. Training Loss over Epochs for CBOW

C. Predictive Modelling

Once features were extracted, various machine learning classifiers are trained to predict whether a tweet is disaster related or not. These models were then compared to see how well they perform on the same dataset. Models included:

- Logistic Regression
- Random Forest
- Gradient Boosting
- SVM
- K-Nearest Neighbors
- Decision Tree

V. Experiments

A. Data Description

The dataset used in this project is a combination of a labeled tweet dataset sourced from Kaggle's *"NLP Getting Started"* competition and additional tweets manually collected by team members.

- **Dataset Characteristics:**
    - Total records: Approximately 12,000 tweets.
    - Labels: Binary classification – *1* for disaster-related tweets and *0* for non-disaster-related tweets.
    - Diversity: Encompasses a wide range of disasters, including natural disasters (e.g., earthquakes, hurricanes) and man-made disasters (e.g., fires, accidents).
    - Preprocessing: Tweets were cleaned by removing URLs, hashtags, mentions, and special characters to standardize the input.

B. Evaluation Metrics

The performance of each model was evaluated using standard classification metrics:

- **Precision:** Precision measures the proportion of correctly predicted disaster tweets out of all tweets predicted as disasters.

Precision = TP / (TP + FP)

- **Recall:** Recall indicates the model's ability to identify disaster tweets from all actual disaster tweets. **Recall = TP /**

(TP + FN)

- **F1-Score:** Harmonic mean of Precision and Recall, balancing false positives and false negatives. **F1-Score =**

2*Precision*Recall / (Precision + Recall)

- **Accuracy:** Percentage of correctly classified tweets across both classes.

Accuracy = TP + FP / (TP + TN + FP + FN)

C. Baseline Methods for Comparison

The following classical machine learning models were employed as baseline methods:

1. **Decision Trees:** Non-linear models that split data into subsets based on feature thresholds.
2. **Random Forest:** An ensemble of decision trees to improve generalization and reduce overfitting. 3. **AdaBoost:** Boosting algorithm that combines weak classifiers to build a strong classifier. 4. **Support Vector Machines (SVM):** Identifies an optimal hyperplane for class separation. 5. **Bayesian Classification:** Probabilistic classifier based on Bayes' theorem.
6. **Logistic Regression:** Linear model for binary classification.
7. **k-Nearest Neighbors (kNN):** Instance-based classifier relying on distance measures.

D. Experimental Setup

The experiments were conducted using three feature extraction methods: TF-IDF, CBOW, and BERT embeddings. Each classifier was trained and tested using the same dataset split (80% training, 20% testing) to ensure consistency.

5.4.1 TF-IDF Experiments

- Logistic Regression emerged as the best-performing model with an F1-score of ~75%. • Decision Trees and kNN showed limited performance due to the sparse nature of TF-IDF features, often leading to overfitting or poor generalization.

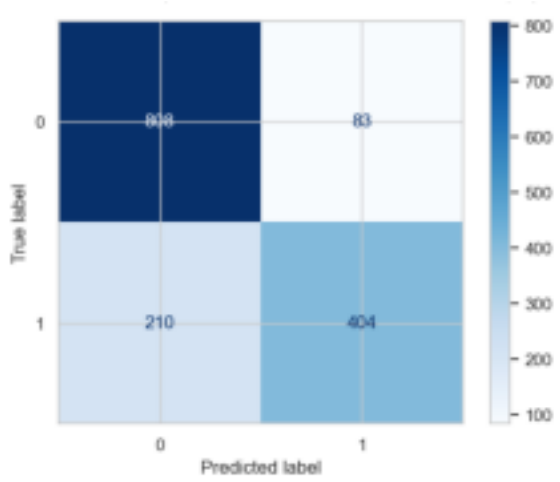

Figure 3. Confusion Matrix for Logistic Regression using TF-IDF as the feature extractor

5.4.2 CBOW Experiments

• CBOW embeddings improved the contextual understanding of tweets.

• SVM outperformed other models with an F1-score of 68%, leveraging the dense embeddings effectively. • Visualizations using t-SNE highlighted clustering patterns, indicating CBOW's ability to capture disaster related tweet contexts.

5.4.3 BERT Experiments

• BERT embeddings significantly enhanced classification performance across all models. • SVM achieved the highest overall F1-score (~78%), validating its compatibility with high-dimensional, dense embeddings.

• Logistic Regression also performed well, with an F1-score of ~74%, while Decision Trees and Bayesian classifiers lagged due to their inability to fully exploit BERT's contextual depth.

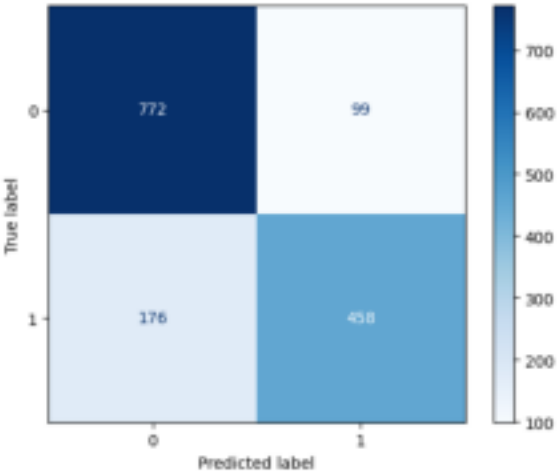

Figure 4. Confusion Matrix for SVM using BERT as the feature extractor

E. Overall Performance Across Metrics

Feature Extraction Type	Best Model	Precision	Recall	F1-Score
TF-IDF	Logistic Regression	0.72	0.77	0.75

CBOW	SVM	0.65	0.70	0.68
BERT	SVM	0.79	0.76	0.81

Table 1. Comparison of different evaluation metrics for the best model of each feature extraction type

VI. Related Work

One of the fundamental methods in natural language processing (NLP), especially in the context of the Word2Vec architecture, is CBOW or continuous bag of words paradigm. Based on the words that surround a target word inside a predetermined window, it is intended to forecast that word. For example, CBOW would use the context terms "The," "university," "at," and "Arizona" to predict "sat" in a sentence like "The university at Arizona." As mentioned in the study [1] this method creates a representation that can anticipate the target word by combining contextual information, usually by averaging embeddings of the surrounding words. The study [2] shows that CBOW, because of its computational simplicity, CBOW is a popular option for large-scale text processing applications because it can be taught more quickly than many alternative models

Additionally, CBOW or continuous bag of words differs from modern transformer-based models such as BERT or Bidirectional Encoder Representation from Transformer. Masked language modeling, a method used by BERT, can be thought of as a more advanced version of CBOW. This method involves masking some words in a sentence and uses their prior and following contexts to forecast the words. Although BERT has a substantially higher computing cost than CBOW, it can capture more semantic nuances due to this bidirectional context. Furthermore, BERT creates
dynamic embeddings that change based on a word's function in a sentence, whereas CBOW embeddings are static, meaning that a word's representation is fixed regardless of its context [3].

Similarly, the study [2] suggests instead of depending on context windows as CBOW does, models like GloVe use co-occurrence statistics throughout the entire corpus. Word analogies and other global semantic links are frequently better captured by GloVe embeddings; nonetheless, they necessitate pre-calculated co-occurrence matrices, which can be memory-intensive for big vocabularies.

A thorough study [4] for comprehending, creating, and training NLP models with BERT or Bidirectional Encoder Representations from Transformers may be found in the book Getting Started with BERT. It is perfect for NLP practitioners and enthusiasts because it covers the model's design, real-world applications, and implementations. The study bridges the gap between theory and practice in state-of-the-art natural language processing by providing step
by-step instructions that allow users to implement BERT for tasks like text classification and sentiment analysis. A version of the BERT model designed especially for Twitter-based sentiment analysis during the COVID-19 pandemic is presented in the publication [5] TM-BERT: A Twitter Modified BERT for Sentiment Analysis on COVID-19 Vaccination Tweets. A domain-specific dataset of tweets pertaining to COVID-19 vaccinations is used to fine-tune TM-BERT. The study shows that TM-BERT is more successful than vanilla BERT, obtaining higher accuracy in tasks involving sentiment classification. It emphasizes the model's capacity to assess public opinion, which is essential for comprehending vaccine skepticism and disinformation.

The shortcomings of models such as K-BERT are addressed by TK-BERT in the study [6], a unique extension of BERT that includes topic-specific knowledge graphs to increase language representation. Although K-BERT integrates domain-specific knowledge graphs, performance may be harmed by the introduction of unrelated knowledge that is not related to the context of the input data. By using a topic modeling technique to divide the knowledge graph into topic-based subsets, TK-BERT gets around this problem and makes sure that only pertinent information is added to the model while it is being processed. Performance on all language tasks improves because of this selective integration, which raises the caliber of knowledge used for training and prediction. In tasks requiring precise contextual awareness, the model outperformed K-BERT, demonstrating its efficacy in

matching language representation with domain-specific issues.

Similarly, the goal of the paper [7] is to apply sentiment classification on social network texts to analyze the emotional tendencies of middle school pupils. To improve accuracy, this study uses ensemble learning to merge many BERT-based models. It balances prediction quality and efficiency by comparing single-layer and deeper BERT networks and applying majority vote among classifiers using Weibo data. Ensemble approaches offer important interpretability for educational mental health assessments, even though deeper BERT networks are recommended for better training results.

The study [8] provides a thorough analysis of text classification strategies designed specifically for the detection of offensive language and hate speech. The study contrasts traditional Globe embeddings incorporated into neural networks like CNN and LSTM with transformer models based on BERT. The findings show that the transformer model for example BERT performs better than GloVe-based techniques in terms of accuracy and resilience, especially when dealing with datasets that are multilingual or domain-shifted. However, because GloVe embeddings
require less computing power, they are still competitive. The results demonstrate the trade-offs between these approaches on several datasets, such as Davidson and Founta.

VII. Conclusion

This research project has demonstrated the significant potential of advanced natural language processing techniques in classifying disaster-related tweets. The comparative analysis of different embedding methods and classification models revealed that BERT-based approaches, particularly when combined with Support Vector Machines, achieve superior performance with an F1-score of 78%. The experimental results highlight a clear progression in classification accuracy from traditional methods to modern approaches. While TF-IDF with Logistic Regression showed respectable performance at 75%, and CBOW with SVM achieved 68%, the BERT-based model emerged as the most effective solution for distinguishing between genuine disaster-related content and metaphorical usage.

This system's practical implications are substantial for disaster response and management. By accurately identifying disaster-related communications in real-time, emergency response teams can better allocate resources and coordinate relief efforts. The model's ability to understand context and nuanced language patterns makes it particularly valuable for real-world applications where quick, accurate decisions are crucial. Future developments could focus on fine-tuning the BERT model for specific types of disasters and incorporating multilingual capabilities to expand the system's global applicability. The success of this project establishes a strong foundation for the continued evolution of automated disaster response systems through social media monitoring.

VIII. References

[1] K. Ganesan, “Word2Vec: A Comparison Between CBOW, SkipGram & SkipGramSI,” *Kavita Ganesan, PhD*, Apr. 22, 2020. https://kavita ganesan.com/comparison-between-cbow-skipgram-subword/

[2] M. Riva, “Word Embeddings: CBOW vs Skip-Gram | Baeldung on Computer Science,” *www.baeldung.com*, Mar. 11, 2021. https://www.baeldung.com/cs/word-embeddings-cbow-vs-skip-gram

[3] Bhoomika Madhukar, “The Continuous Bag Of Words (CBOW) Model in NLP - Hands-On,” *Analytics India Magazine*, Sep. 10, 2020. https://analyticsindiamag.com/ai-mysteries/the-continuous-bag-of-words-cbow-model-in-nlp-hands-on-implementation-with-codes/ (accessed Nov. 24, 2024).

[4] Sudharsan Ravichandiran, *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT* , Packt Publishing, 2021.

[5] M. T. Riaz, M. Shah Jahan, S. G. Khawaja, A. Shaukat and J. Zeb, "TM-BERT: A Twitter Modified BERT for Sentiment Analysis on Covid 19 Vaccination Tweets," *2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2)*, Rawalpindi, Pakistan, 2022, pp. 1-6, doi: 10.1109/ICoDT255437.2022.9787395.

[6] C. Min, J. Ahn, T. Lee and D. -H. Im, "TK-BERT: Effective Model of Language Representation using Topic-based Knowledge Graphs," *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Seoul, Korea, Republic of, 2023, pp. 1-4, doi: 10.1109/IMCOM56909.2023.10035573.

[7] K. Jiang, H. Yang, Y. Wang, Q. Chen and Y. Luo, "Ensemble BERT: A Student Social Network Text Sentiment Classification Model Based on Ensemble Learning and BERT Architecture," *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, Jinzhou, China, 2024, pp. 359-362, doi: 10.1109/ICSECE61636.2024.10729490.

[8] S. S, U. S, N. Abinaya, J. P, S. Priyanka and D. M N, "A Comparative Exploration in Text Classification for Hate Speech and Offensive Language Detection Using BERT-Based and GloVeEmbeddings," *2024 2nd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2024, pp. 1506-1509, doi: 10.1109/ICDT61202.2024.10489019.

[9] D. Kothadiya, C. Bhatt, K. Sapariya, K. Patel, A. B. Gil-González, and J. M. Corchado, "Deepsign: Sign Language Detection and Recognition Using Deep Learning," Electronics (Switzerland), vol. 11, no. 11, Jun. 2022, doi: 10.3390/electronics11111780

[10] S. Dhulipala, F. F. Adedoyin, and A. Bruno, "Sign and Human Action Detection Using Deep Learning," J Imaging, vol. 8, no. 7, Jul. 2022, doi: 10.3390/jimaging8070192.