Journal of Data Applications
Issue: 3, 2024
E-ISSN: 2980-3357

İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

Detection of Urgent Messages Shared on Twitter during an Earthquake using the Deep Learning Method

Mücahit SÖYLEMEZ[1], Ali ÖZTÜRK[2,3]

ABSTRACT

The ability of Twitter to provide real-time information during disasters is becoming more widely acknowledged, making it an essential forum for people to voice their concerns and ask for help during emergencies. These platforms can speed up the distribution of help, but they are also prone to false information, which might make disaster response more difficult. Using a carefully selected dataset of 10,200 tweets that have been extensively preprocessed and tokenized for reliable training and validation, this study uses deep learning models, such as LSTM, BLSTM, and BLSTMA, to classify tweets during earthquake events into two categories: "under the debris" and "not under the debris." The model performance was further improved via hyperparameter adjustment, which included neuron counts, dropout rates, dimensions, and embedding types. The results of this study showed that while the BLSTMA model had the best accuracy (96.64%) and F1 score (0.9116), conventional machine learning techniques like XGBoost and SVM. However, in other measurements, it was shown that standard machine learning techniques like SVM and XGBoost performed better. Using Bag of Words vectorisation, SVM obtained 95.81% accuracy and an F1 score of 0.9579, whereas XGBoost earned 95.84% accuracy and an F1 score of 0.9584. By demonstrating the usefulness of the BLSTMA model in real-time disaster response and the complementary advantages of conventional approaches in the analysis of complex disaster data, these findings highlight the significance of customising machine learning and deep learning approaches to particular tasks.

Keywords: Deep Learning, Natural Language Processing, Disaster Management, Twitter Analysis, Emergency Message Detection.

DOI: 10.26650/JODA.1606403

[1]İstanbul University, Faculty of Economics, Department of Management Information Systems, İstanbul, Türkiye
[2]KTO Karatay University, Department of Computer Engineering, Konya, Türkiye
[3]Chief Systems Engineers, Havelsan A.Ş., Ankara, Türkiye

ORCID: M.S. 0009-0009-0059-616X;
A.Ö. 0000-0002-1797-2039

Corresponding author:
Mücahit SÖYLEMEZ,
İstanbul University, Faculty of Economics, Department of Management Information Systems, İstanbul, Türkiye
E-mail: mucahit.soylemez@istanbul.edu.tr

Submitted: 24.12.2024
Revision Requested: 16.01.2025
Last Revision Received: 21.01.2025
Accepted: 22.01.2025
Published Online: 31.01.2025

Citation: Söylemez, M., & Öztürk, A. (2024). Detection of Urgent Messages Shared on Twitter during an Earthquake using the Deep Learning Method. *Journal of Data Applications, 3*, 63-84. https://doi.org/10.26650/JODA.1606403

Introduction

Natural disasters are among the foremost events that threaten human life and economies on a global scale. Among these, earthquakes stand out as the most significant natural disaster, leading to the loss of human life. The necessity for countries on active fault lines to establish an effective disaster management system becomes increasingly evident with each passing day. Therefore, leveraging all available technological resources should be a primary objective for these nations. The greater the destruction caused by a disaster; the more critical communication becomes. Considering this, implementing additional measures in disaster management has become an essential requirement for every country. For instance, Hurricane Harvey, which struck the United States in certain years, caused significant hardship for millions of Americans, prompting the U.S. to take additional precautions. Similarly, taking extra measures against earthquakes, which are even more impactful, should be a priority in the disaster management strategies of all countries.

The role of social media platforms in crisis management has gained increasing importance with advancements in technology. Platforms like Twitter, in particular, have become critical tools for communication during disasters, facilitating rapid information flow for both individuals and aid teams. In recent years, during an earthquake in Turkey, survivors trapped under debris resorted to sharing their locations and conditions through social media platforms like Twitter due to the inability to communicate via traditional phone lines or conventional communication channels (Al Jazeera, 2023). These platforms proved instrumental for rescue teams, helping to save many lives (Euronews, 2023). However, the dissemination of false information on such social media platforms could not be entirely prevented.

Earthquakes, as one of the most destructive natural disasters, lead to significant loss of life, widespread damage to infrastructure, and disruption of daily life. In the aftermath of an earthquake, many individuals are trapped under debris, and the timely identification of those in need of rescue becomes critical. In such chaotic situations, traditional communication systems may fail, and the flow of information can be severely hindered, which delays rescue efforts and worsens the crisis.

Social media platforms, particularly Twitter, play a pivotal role in these scenarios by providing individuals with the opportunity to share real-time information about their location, conditions, and urgent needs (Al Jazeera, 2023). While this communication channel holds the potential to significantly aid disaster response teams, it also comes with the challenge of filtering out irrelevant or false information. To address these challenges, this study aims to analyse tweets shared by individuals trapped under debris during earthquakes using a deep learning-based artificial intelligence system. The goal is to accurately determine whether these individuals are truly in need of rescue, thereby enhancing the effectiveness of disaster management efforts by providing reliable information (Euronews, 2023).

Literature Review

Social media platforms have become a crucial source of information for crisis management. In this context, Powers et al. (2023) investigated how social media may help during natural disasters by detecting emergency signals. With an emphasis on tweets from Hurricane Harvey, they investigated many methods for recognising messages from persons in imminent need. Their results showed that certain models, such as XLNet and BERT, outperformed others; CNN's accuracy was 72%, but BERT's was 78%. The paper makes the case that more data might further improve these systems' efficacy while highlighting the importance of social media in disaster response. (Powers et al., 2023)

During natural disasters, Pradip Bhere and colleagues investigated the use of deep learning to differentiate between relevant and irrelevant tweets. To improve tweet classification, they created a system that combined a convolutional neural network (CNN) with Word2Vec feature vectors. They tested the model with different word embeddings (Custom Weight, Google News, Twitter Glove) using a dataset of 10,000 tweets from Kaggle that dealt with disasters. According to the results, the model's accuracy using Google News embeddings was 86%, while its accuracy using their proposed method was 84%. In addition, they evaluated its accuracy using a confusion matrix on tweets from the 2013 Colorado floods. This work demonstrates how deep learning can be used to identify social media data during emergencies and raises the possibility that future performance could be improved with larger datasets. (Bhere et al., 2020).

Kumar et al. compared several machine learning and deep learning techniques for classifying social media tweets regarding catastrophes in order to examine performance under data imbalance. The study found that the deep learning models performed better than the conventional classifiers. For the hurricane dataset, BIGRU got the greatest F1 score (0.87), while for the earthquake dataset, GRU-CNN had the highest F1 score (0.88). These findings demonstrate how well deep learning models categorise tweets concerning catastrophes and how they may improve local disaster response activities. (Kumar et al., 2019).

Behl et al. (2021) investigated a range of machine learning and deep learning models to classify Twitter data during COVID-19 and natural catastrophe occurrences. They examined various models, including multilayer perceptions (MLP-TF and MLP-W), convolutional neural networks (CNN-W and CNN-WF), and logistic regression (LR-TF). Their findings revealed that the LR-TF model achieved the best performance, with 88% accuracy on a dataset of earthquakes in Nepal and Italy and 81% accuracy on the COVID-19 dataset. Both CNN-W and CNN-WF delivered similar performance across the datasets, though CNN-WF's accuracy was slightly lower at 78% on the COVID-19 dataset. The accuracy of MLP-TF decreased from 87% on the combined dataset to 77% on the COVID-19 dataset. In contrast, the MLP-W model performed best on the COVID-19 dataset, achieving an accuracy of 83% (Behl et al., 2021).

Madichetty et al. developed a Stacked Convolutional Neural Network (SCNN) model to identify resource-related tweets during emergencies. The model integrates CNN and KNN classifiers at the base level, with an SVM meta-classifier processing the outputs for the final classification. When tested on datasets from the 2015 Nepal and 2016 Italy earthquakes, the SCNN model outperformed other combinations, achieving the highest accuracy of 77.5% for Nepal and 76.99% for Italy. These results highlight the effective collaboration of CNN, KNN, and SVM in categorizing social media data for disaster management (Madichetty & Sridevi, 2020).

Muhammed Ali Sit and colleagues analysed tweets during Hurricane Irma to explore the use of social media in disaster management. They found that Long Short-Term Memory (LSTM) networks performed best, achieving 74.78% accuracy and 75.14% F1 score in classifying disaster-related tweets. Other methods, such as CNN and logistic regression, performed less well. The study highlights LSTM as the most effective model for analysing social media data during crises.

All these studies have highlighted the potential of social media platforms in crisis management and have provided valuable contributions to the disaster management literature using deep learning, machine learning, and natural language processing techniques.

In our study, tweets from individuals trapped under the debris were successfully identified using natural language processing and deep learning models applied to a unique dataset.

Studies in the literature emphasise the importance of artificial intelligence-based approaches developed for analysing and interpreting social media data during crises. Various research efforts have introduced innovative methods using natural language processing, deep learning, and machine learning techniques to contribute to disaster management processes. Below is a summary of the contributions these studies have made to the literature.

Materials and Methods

Artificial intelligence, with a deep-rooted history spanning from ancient times to the modern era, has evolved into a discipline that finds applications in almost every field of contemporary technology (Mijwel, 2015). This study focuses on deep learning and natural language processing, which are subfields of artificial intelligence, aiming to effectively classify social media data for disaster management. To achieve this goal, the performance of the models developed using deep learning techniques and natural language processing methods was evaluated. We implemented a machine learning pipeline to classify the textual data using various recurrent neural network (RNN) architectures. The preprocessing steps included converting text to lowercase, removing URLs and special characters, tokenization, stopwords removal, and stemming using the TurkishStemmer library. The dataset was split into training and validation sets, and word embeddings were generated using both a pre-trained Word2Vec model and random initialisation. Three RNN architectures—LSTM, bidirectional LSTM (BLSTM), and BLSTM with attention (BLSTMA)—were constructed using TensorFlow and Keras. Each architecture was configured with different hyperparameters, such as the number of units, dropout rates, and learning rates, to evaluate their performance. Training was conducted using class-balanced weights and monitored using callbacks for early stopping and learning rate adjustment. Performance metrics, including accuracy, precision, recall, and F1 score, were calculated to assess the effectiveness of each model configuration. The experimental results are presented in detail in the subsequent sections.

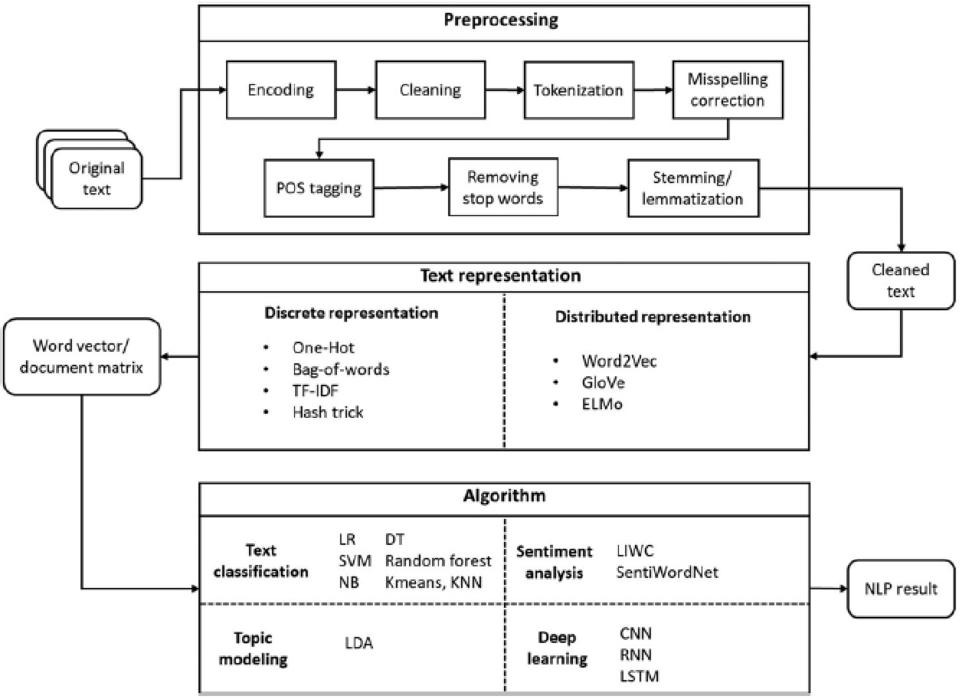

Figure 1. *Natural Language Processing Steps (Kang et al., 2020)*

Natural Language Processing (NLP)

Natural Language Processing (NLP) is an area of artificial intelligence that focuses on enabling machines to understand, interpret and respond to human language in a meaningful way. The core of NLP is the transformation of raw textual data into structured forms suitable for machine learning algorithms. This transformation involves a series of preprocessing steps, text representation techniques, and algorithmic methods tailored to specific applications such as sentiment analysis, text classification, and machine translation (Kang et al., 2020).

Preprocessing in the NLP

A number of preprocessing steps are taken at the beginning of the NLP process to prepare the textual input for the machine learning algorithms. These steps include encoding compatibility, cleaning up HTML elements and redundant components, segmenting text into units (tokenisation), spelling correction, tagging word types (POS tagging), removing stop words, and reducing words to base or root forms (stemming or lemmatisation). These procedures ensure that the raw text is standardised and cleaned, allowing for better performance of downstream NLP models (Kang et al., 2020). Figure 1 shows these steps in detail. The effects of these preprocessing steps are discussed in the Dataset and Experimental Results sections.

Text representation techniques

Once pre-processing is complete, the cleaned text is converted into numerical representations for the machine learning models. This conversion can be performed using separate or distributed representation methods:

- Discrete representations: Techniques such as One-Hot Encoding, Bag-of-Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF) and the shing trick are used to represent text as vectors based on word occurrence or importance within a document. These methods capture basic textual information but may lose contextual relationships (Kang et al., 2020).

- Distributed representations: Advanced methods such as Word2Vec, GloVe, and FastText generate dense vector representations that preserve the semantic relationships between words. These embeddings are particularly useful for capturing word meaning and context, making them suitable for tasks such as sentiment analysis and machine translation (Kang et al., 2020).

Applications such as text categorisation, machine translation and sentiment analysis often use these representation techniques. Discrete representations are often paired with traditional machine learning algorithms such as logistic regression (LR), support vector machines (SVM), and decision trees (DT), while distributed representations are often integrated with neural networks such as convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) (Kang et al., 2020).

This study was undertaken for evaluating the effectiveness of various NLP techniques in a text classification task. As outlined in Figure 1, the NLP process begins with preprocessing, the aim of which is to clean and standardise the raw text. Subsequently, both discrete and distributed text representation techniques were applied to convert the text into machine-readable formats. Finally, these representations were used to train and evaluate classification models, such as LSTM-based neural networks. The study's objective is to determine the most effective techniques for enhancing text classification accuracy. In summary, this study underscores the significance of preprocessing, investigates diverse text representation methods, and employs sophisticated algorithms to attain optimal performance in NLP tasks. Figure 1 offers a comprehensive visual depiction of the entire NLP pipeline (Kang et al., 2020).

Dataset

For the analysis of messages shared on Twitter within the scope of disaster management, a dataset comprising 10,200 tweets was prepared for the period between February 6 and 8. The data collection process was conducted using the Python programming language and the SntTwitter library. During this process, tweets were collected within a specific focus area by using disaster-related hashtags (e.g., #earthquake, #help). The collected data were recorded with attributes such as date, content, and hashtags used and subsequently manually classified.

The classification process divided the tweets into two categories: "emergency help messages" (1) and "general information sharing" (0). This classification was performed to facilitate the training process of the supervised learning algorithms. Table 1 provides examples from the dataset, offering a framework for understanding how social media data are processed in the context of disaster management.

Table 1. *Twitter Dataset*

Date	Tweet	HashTags	Label
2023-02-06 23:59:56+00:00	SİTE 1 NO:20 HATAY MERKEZ MELİSA YARDIM BEKLİYOR	['afaddeprem','yardım','deprem' 'afad']	1
2023-02-06 29:59:51+00:00	Aksever Mahallesi Meltem Sokak Güler Apartmanı	['hatayyardımbekliyor','hatayafad', 'hatayardım']	1
2023-02-06 23:59:38+00:00	ÖNEMLİ DUYURU YAYALIM hatay hatayyardımbekliyor	['hatay','hatayyardımbekliyor', 'ENKAZALTINDA'	0

To provide a general perspective by analysing the frequency and contexts of words within the dataset, a word cloud analysis was conducted. Figure 2 visualises the most frequently used words, highlighting key themes that are prominent in the context of disaster management. This analysis contributes significantly to the identification of critical themes and patterns in the data.

Figure 2. *Word Cloud in the Dataset*

The dataset predominantly features earthquake-related emergency tweets, where the most frequent words—"deprem" (earthquake), "yardım" (help), "lütfen" (please), and "enkaz" (debris)—indicate a high volume of distress calls and urgent requests for rescue operations. As illustrated in Figure 3, the frequency distribution of these keywords highlights their critical role in disaster-related communication. Additionally, the presence of geographical terms like "Hatay" and references to AFAD (Turkey's Disaster and Emergency Management Authority) suggests a focus on specific affected regions and official response efforts, emphasising the critical need for immediate aid and coordinated rescue missions.

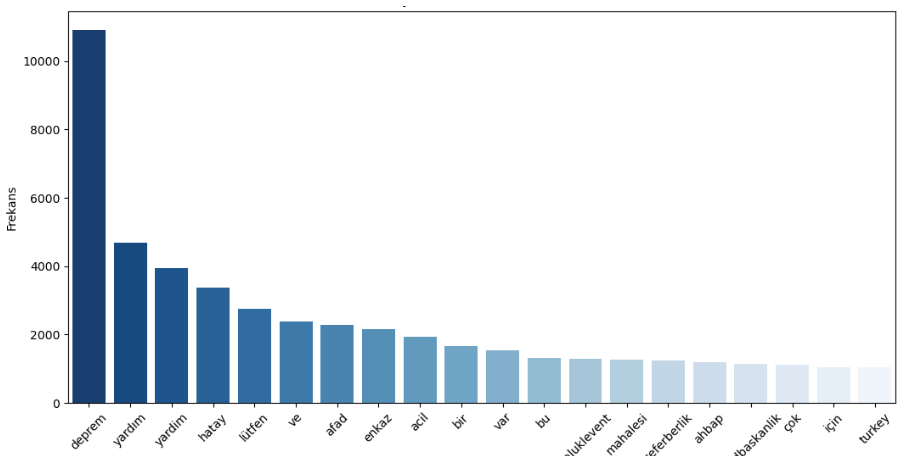

Figure 3. *Frequency analysis of the 20 most frequently mentioned words*

Deep Learning Methods

Deep learning is a branch of machine learning that provides effective solutions to complex problems by extracting meaningful insights from large datasets (Goodfellow, Bengio, & Courville, 2016). This method, particularly prominent in sequential data, natural language processing, and time-series analysis, has enabled groundbreaking advancements in data analysis processes. At the core of deep learning techniques lie artificial neural networks (ANNs), which are inspired by the neural system of the human brain. ANNs are structures capable of solving non-linear problems by learning input and output values and processing them through a specific algorithm to generate results (Goodfellow, Bengio, & Courville, 2016). The development of this field, which began with single-layer perceptions, led to the creation of multilayer perceptions (MLPs) to address the need for understanding non-linear relationships and complex data structures. MLPs, capable of learning more intricate relationships, have found wide applications in classification and regression problems within machine learning (Rosenblatt, 1958; Goodfellow, Bengio, & Courville, 2016). However, the processing of sequential and time-dependent data highlighted the limitations of ANNs, prompting the development of Recurrent Neural Networks (RNNs) and their derivatives.

Recurrent Neural Networks (RNNs) are specialised architectures designed to work with sequential data and are capable of learning from sequential inputs (Goodfellow, Bengio, & Courville, 2016). However, RNNs face challenges such as gradient vanishing when learning long-term dependencies. To address these issues, **Long Short-Term Memory (LSTM)** models were developed. LSTM models employ mechanisms such as forget, input, and output gates, effectively controlling the flow of information and excelling in learning long-term dependencies (Hochreiter & Schmidhuber, 1997). Due to these features, LSTM models are widely used in fields such as text processing, time-series analysis, and natural language processing (Hochreiter & Schmidhuber, 1997; Graves, Mohamed, & Hinton, 2013).

An advanced version of LSTM, **Bidirectional Long Short-Term Memory (BLSTM),** enhances contextual representation by learning both past and future contexts in sequential data (Zhou et al., 2016). BLSTM processes information bidirectionally, achieving robust results in preserving semantic coherence, particularly in text data. The performance of the BLSTM further improves when combined with the **attention mechanism**. The attention mechanism enables the model to focus on critical inputs, prioritising key information, especially in long and complex sequences (Zhou et al., 2016).

Finally, the **Bidirectional Long Short-Term Memory with Attention (BLSTMA)** model, which integrates the attention mechanism into BLSTM, not only remembers information during the learning process but also focuses on the most critical elements, achieving higher accuracy and efficiency (Zhong et al., 2020). This progressive evolution of the LSTM, BLSTM, and BLSTMA models significantly enhances the sequential data processing capabilities of artificial neural networks, paving the way for groundbreaking applications in various fields, particularly natural language processing.

Training Parameters Of Deep Learning Models

In this study, a comprehensive performance comparison of the deep learning algorithms was conducted using metrics such as accuracy, precision, recall, and F1 scores. In addition, various combinations of hyperparameters were tested, including random embedding and FastText-based embedding techniques with representations of 100, 200, and 300 dimensions; neuron counts of 64, 128, and 256 units; and dropout rates of 0.2, 0.3, and 0.4. The results were evaluated to assess the potential of optimising disaster management tasks, such as analysing social media data, using deep learning algorithms (LSTM, Bidirectional LSTM - BLSTM, and Attention-augmented Bidirectional LSTM - BLSTMA).

A total of 154 variations were created by testing these hyperparameter combinations within the framework of the four specified parameter sets. For each variation, the average accuracy values were calculated to enable a comprehensive performance comparison of the models, particularly in their application to disaster management. This analysis identified the optimal configuration for effectively processing and interpreting social media data in crisis scenarios.

Experimental Results

The dataset used in this study contained unnecessary words and was cleaned by applying the data preprocessing steps outlined in Figure 1. The data cleaning steps employed in this context can be summarised as follows:

- Turkish stopword removal: Common but semantically insignificant words such as “ve” (and) and “bir” (a) were removed from the dataset using the NLTK library.

- Removal of URLs and username tags: The URLs and username tags present in the tweets were extracted from the dataset.

- Cleaning of special characters: Emojis and special characters such as “#” and “$” found in the tweet text were removed.

- Standardisation of uppercase and lowercase letters: All text was converted to lowercase to ensure consistency in character usage.

- Elimination of repeated letters: Expressions with unnecessarily repeated letters were corrected.

These preprocessing steps were performed to make the data more meaningful and analysable. Once the preprocessing was completed, the tweets were tokenised by splitting them into word sequences. During the tokenization step, each word was assigned a unique integer value. For a detailed explanation of the data preprocessing and tokenization steps, refer to Table 2.

Table 2. *Tokenization Process*

Step	Result
Raw Text	SİTE 1 NO:20 HATAY MERKEZ MELİSA YARDIM BEKLİYOR afaddeprem yardım deprem afad
Convert to Lowercase	site 1 no:20 hatay merkez melisa yardım bekliyor afaddeprem yardım deprem afad
Remove Punctuation	site 1 no20 hatay merkez melisa yardım bekliyor afaddeprem yardım deprem afad
Tokenize Words	[‘site’,’1’,’no20’,’hatay’,’merkez’,’melisa’,’yardım’,’bekliyor’,’afaddeprem’, ‘yardım’,’deprem’,’afad’]
Assign Unique Index	{‘site’:1,’1’:2,’no20’: 3, ‘hatay’: 4, ‘merkez’: 5, ‘melisa’: 6, ‘yardım’: 7, ‘bekliyor’: 8, ‘afaddeprem’: 9, ‘yardım’: 10, ‘deprem’: 11, ‘afad’: 12}
Convert Words to Indexes	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
Fixed-Length Sequence	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]

In the deep learning approach, the dataset was divided into training and validation sets, with the training set used for model learning and the validation set for evaluating model performance. The models were trained using the binary_crossentropy loss function and the Adam optimisation algorithm. To prevent overfitting and enhance performance, the early stopping method was applied.

The model was trained for 30 epochs, with the data split into training, validation, and test sets in a 67%-33% ratio. During training, early stopping was applied by monitoring the val_loss metric, ensuring that the best weights were retained to avoid performance degradation. Additionally, data augmentation techniques such as word order shuffling and random word dropping were used to increase the model's generalisation capacity. The model was trained using Word2Vec and random embedding matrices and structured with LSTM, Bidirectional LSTM (BLSTM), and Attention-based BLSTM (BLSTMA) architectures. The models were trained with class weights, continuously monitored with validation data, and the model with the highest accuracy was selected.

During the training process, the classification models learned the context of the tweets, enabling them to accurately classify new messages. The changes in accuracy based on the number of epochs are illustrated in Figures 4, 5, and 6.

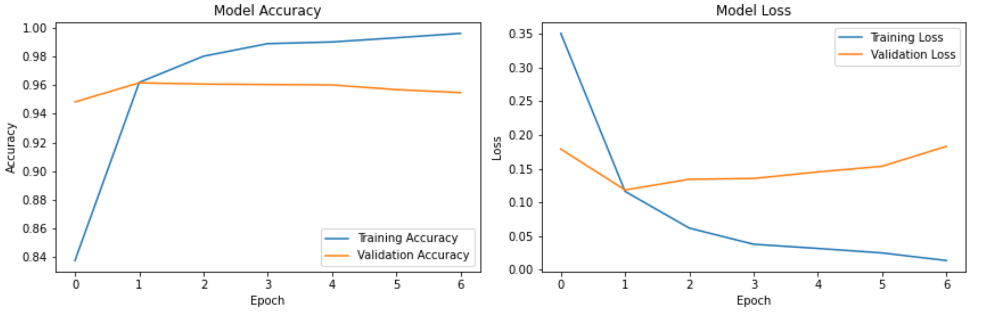

Figure 4. *LSTM Training Process*

As seen in Figure 4, the graphs show the LSTM model's training and validation performance throughout the six epochs. It is clear that by the fourth epoch, the training accuracy has increased consistently to about 98%, while the validation accuracy has stabilised at a somewhat lower level, between 94% and 95%. Additionally, after the second epochs, the training loss steadily dropped from 0.9 to about 0.05, while the validation loss reached a plateau at 0.25. Nevertheless, the difference between these two forms of loss remains, indicating that the LSTM model has trouble generalising, most likely because it is inadequate at capturing long-range dependencies (Hochreiter & Schmidhuber, 1997).

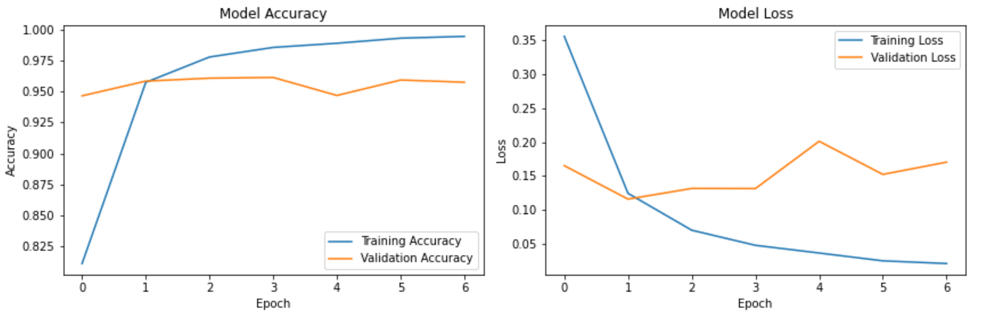

Figure 5. *BLSTM Training Process*

The BLSTM model has 99% training accuracy in the first two epochs, as shown in Figure 5. When compared to the LSTM model, the validation accuracy shows a little improvement, consistently hitting a 95% level. After the second epoch, the validation loss converges at about 0.2. The initial training loss of the BLSTM model is 0.8, and it eventually drops to less than 0.05. In conclusion, the smaller gap between training and validation loss indicates that the BLSTM model's bidirectional nature is better than the unidirectional LSTM model in capturing contextual information (Schuster & Paliwal, 1997).

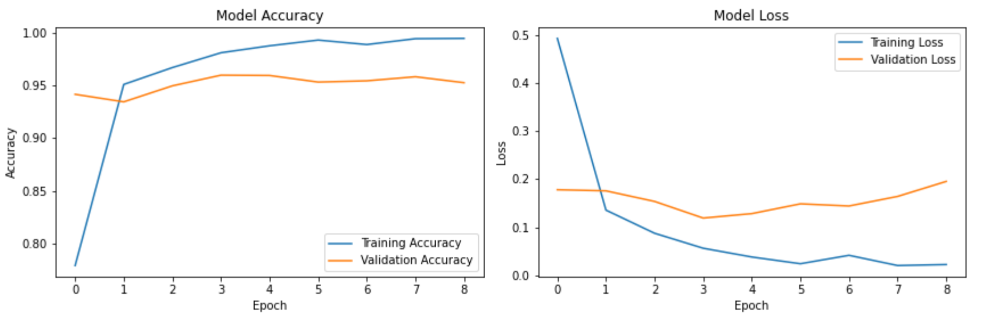

Figure 6. *BLSTMA Training Process*

As demonstrated in Figure 6, the performance of the BLSTM with Attention (BLSTMA) model across eight epochs indicates its effective learning and generalisation capabilities. The training accuracy (blue line) rapidly attains close to 99% by the second epoch and stabilises, while the validation accuracy (orange line) progressively increases, reaching approximately 96%, with a negligible discrepancy between the two. The training loss (BLUE line) displays a sharp decrease from 0.5 to around 0.02 by the eighth epoch, whereas the validation loss (ORANGE line) stabilises at approximately 0.18 (Vaswani et al., 2017). The narrow gap between the training and validation metrics reflects the BLSTMA model's ability to effectively generalise without overfitting. This ability is attributed to the inclusion of attention mechanisms, which enhance the model's focus on significant input features, as discussed in the following section. (Vaswani et al., 2017).

The confusion matrix of the BLSTMA model, which achieved the best results in the deep learning models, is shown in Figure 7. This shows that the model correctly classified 2670 negative samples (true negatives) and 583 positive samples (true positives). There were 60 false positives (negative samples misclassified as positive) and 53 false negatives (positive samples misclassified as negative). These results highlight the model's strong ability to handle imbalanced data with high precision and recall for both classes.

The differences in performance between the BLSTMA and other models likely stem from the attention mechanism, which enables the model to focus on the most relevant features of the input data. This targeted focus improves the model's capacity to capture subtle patterns, especially in complex datasets. Additionally, the use of FastText embeddings contributes to better word representation, capturing semantic and syntactic nuances. In comparison, models without attention mechanisms may struggle to distinguish between similar samples, leading to slightly higher misclassification rates.

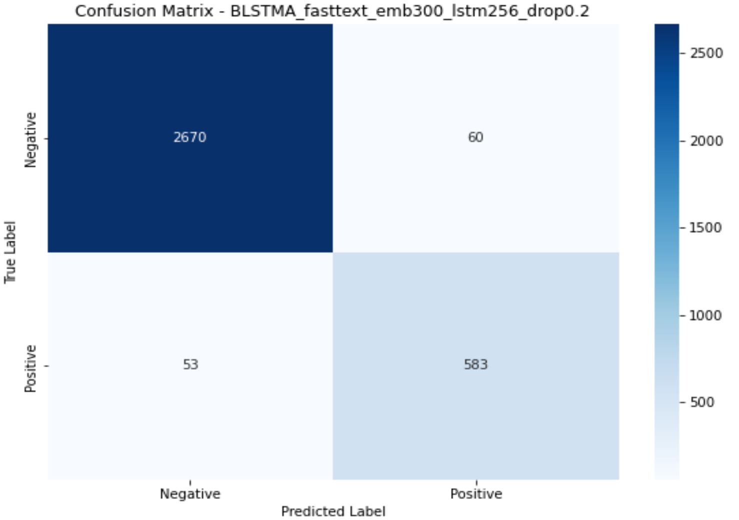

Figure 7. *Confusion Matrix of the BLSTMA Model*

The performance metrics of the deep learning models obtained with the best-performing parameters are presented in Table 3.

Table 3. *Performance comparison of the deep learning models*

Model Type	Embedding Type	Embedding Size	LSTM Unit Count	Dropout Rate	Accuracy	Precision	Recall	F1 Score
LSTM	Random	300	128	0.4	0.964646	0.889894	0.927673	0.908391
BLSTM	Random	300	256	0.4	0.965835	0.895296	0.927673	0.911197
BLSTMA	Fasttext	300	256	0.2	0.966429	0.906687	0.916667	0.911650

In this study, where approximately 162 different parameters were evaluated, the best results are presented in Table 3. To better analyse the performance of the different models, the average accuracy values were calculated based on four selected parameters and compared across the models.

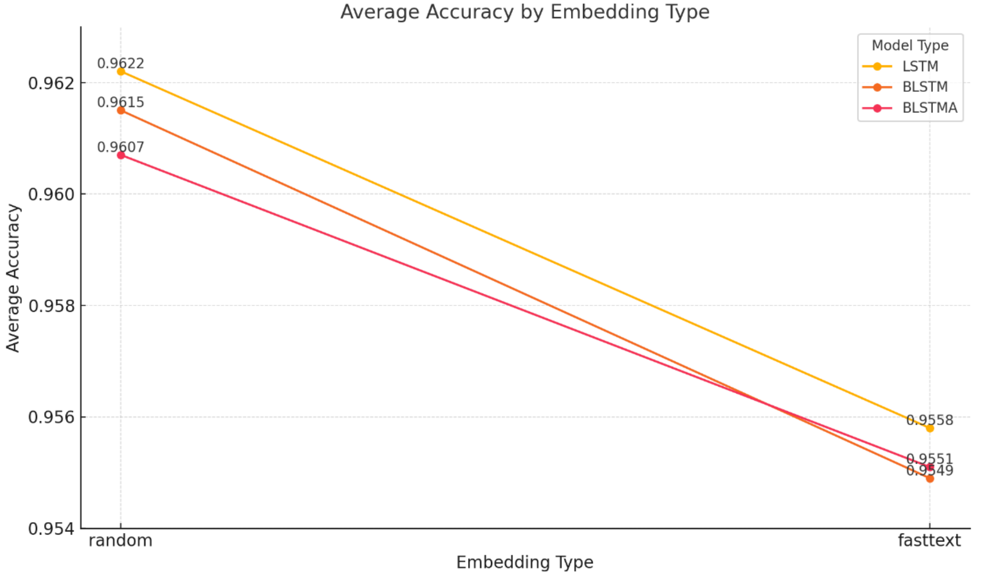

Figure 8. *Average accuracy values for embedding types*

In Figure 8, the BLSTMA model demonstrated the lowest average accuracy with the random embedding type but outperformed the BLSTM model when using the FastText embedding type. This indicates that BLSTMA, while struggling to learn meaningful representations from randomly initialized embeddings, benefits significantly more from pre-trained embeddings compared to BLSTM.

The LSTM model achieved the highest average accuracy of 96.22% with the random embedding type, surpassing the other models. This result indicates that LSTM is more effective at learning contextual relationships without relying on pre-trained embeddings. One possible reason is that LSTM's simpler structure, compared to bidirectional architectures such as BLSTM and BLSTMA, allows it to generalise better when embeddings are not pre-trained.

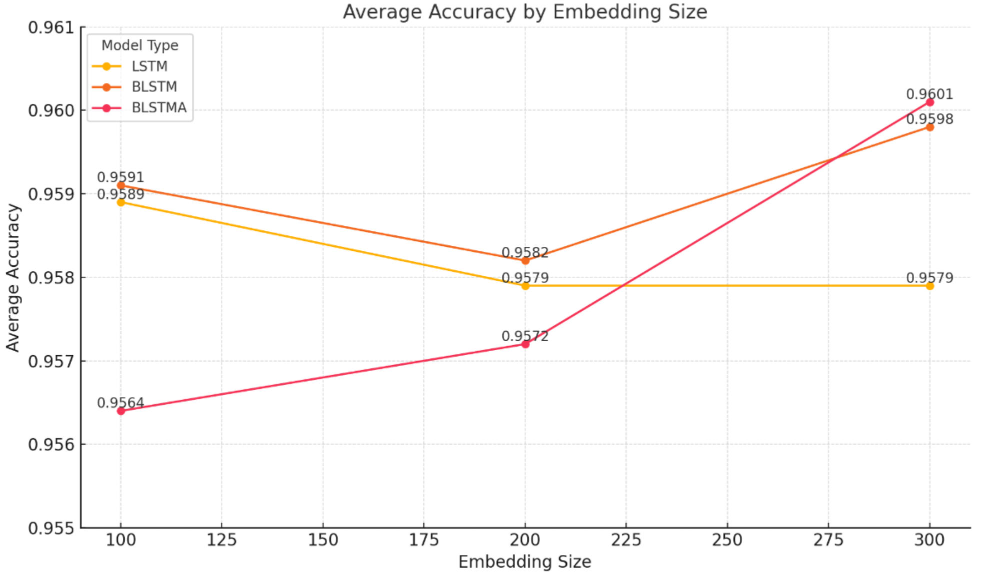

Figure 9. *Average accuracy values for embedding dimensions*

In Figure 9, the BLSTMA model exhibited lower accuracy rates compared to other models with smaller embedding dimensions. However, when the embedding dimension was set to 300, it surpassed the average accuracy values of the LSTM and BLSTM models. This result indicates that the BLSTMA model benefits more from larger embedding sizes due to its bidirectional structure, which can leverage richer feature representations when more parameters are available.

The LSTM model maintained a relatively stable performance as the embedding dimension increased, indicating that it does not significantly benefit from higher-dimensional word embeddings. Meanwhile, the BLSTM model showed an increasing trend, aligning with findings in previous studies that suggest that bidirectional architectures perform better with larger embedding sizes due to their ability to capture forward and backward dependencies more effectively.

The results show that the embedding dimension significantly affects how well the BLSTM and BLSTMA models provide accuracy. However, it is important to remember that overfitting may result from an overabundance of the embedding dimension augmentation. To evaluate their effect on model performance, the embedding dimensions of 100, 200, and 300 were carefully chosen for this work based on experimental considerations.

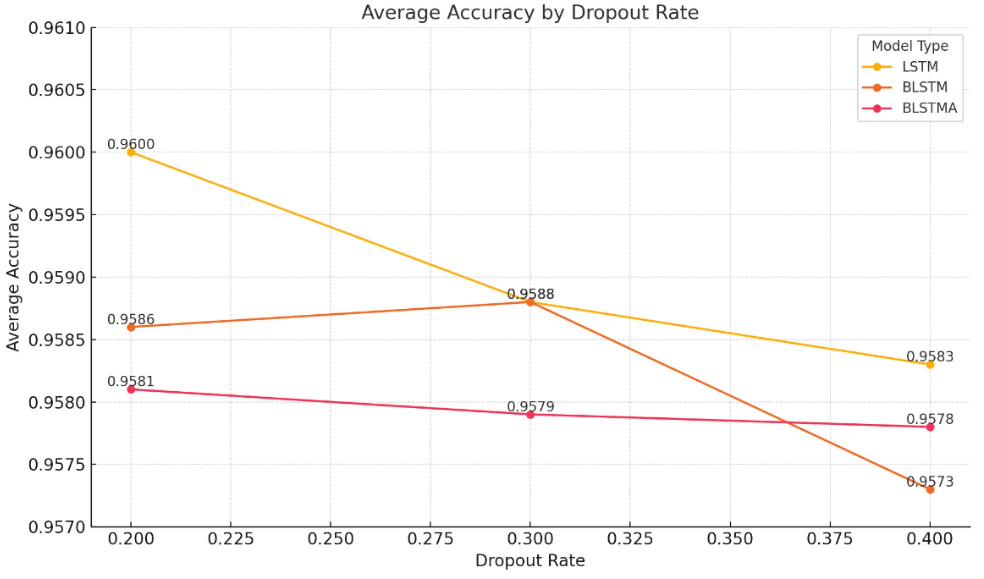

Figure 10. *Average accuracy values for dropout rates*

For the dropout rate parameter, values of 0.2, 0.3, and 0.4 were used. While the LSTM model exhibited a noticeable decline in accuracy beyond a dropout rate of 0.3, both the BLSTM and BLSTMA models began to show reductions in accuracy starting from 0.2. This indicates that the LSTM model experienced significant information loss after 0.3, whereas the BLSTM and BLSTMA models began losing information at 0.3.

The optimal parameter values for these models correspond to the points immediately before the decline in performance begins. As seen in Figure 10, the LSTM model fell below the BLSTMA model in accuracy at a dropout rate of 0.4. Among all the parameters, the BLSTM model consistently provided the best results. This demonstrates that the BLSTM model achieves better generalisation and exhibits a more balanced performance compared to the other models.

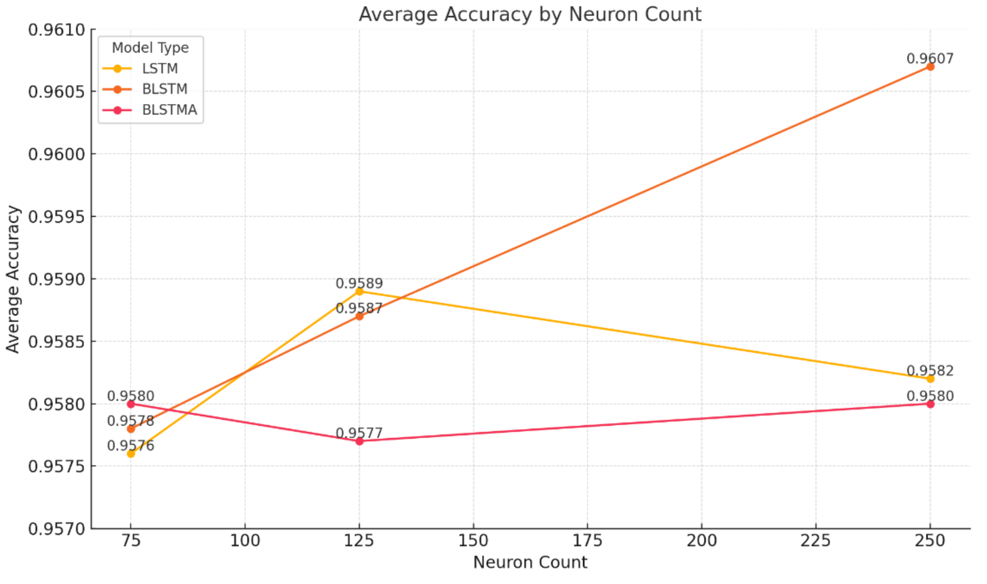

Figure 11. *Average accuracy values for the number of neurons*

As demonstrated in Figure 11, there is a clear indication of how the neuron count affects the performance of the LSTM, BLSTM and BLSTMA models. The BLSTM model's consistent enhancement in accuracy with rising neuron numbers can be ascribed to its capacity to effectively capture the bidirectional context. As the number of neurons escalates, the model evidently gains from the augmented capacity to depict intricate patterns, culminating in superior performance compared to the LSTM and BLSTMA models This indicates that the bidirectional structure in the BLSTM scales effectively with more computational resources. In contrast, the BLSTMA model demonstrates less dependency on the neuron count, achieving competitive accuracy with fewer neurons but failing to leverage additional neurons to the same extent as BLSTM. This could indicate that the attention mechanism primarily enhances local feature learning rather than relying heavily on the increased model capacity. The findings further indicate that the LSTM model demonstrates a decline in accuracy beyond 128 neurons, which may suggest that it experiences overfitting or diminishing returns when scaled due to its lack of bidirectional or attention-enhanced mechanisms to effectively utilise the additional neurons.

The observed differences between these algorithms underscore the merits of the bidirectional context in BLSTM in leveraging higher neuron counts, while the BLSTMA model's attention mechanism underscores efficiency with fewer resources. These observations underscore the necessity for selecting a model architecture that considers not only the characteristics of the data set but also the available computational resources.

Table 4. *Machine Learning Models Results*

Model Type	Vectorisation Method	Accuracy	Precision	Recall	F1 Score
Logistic Regression	TF-IDF	0.9465	0.9454	0.9465	0.9453
Naïve Bayes	TF-IDF	0.9435	0.9438	0.9435	0.9436
SVM	TF-IDF	0.9530	0.9528	0.9530	0.9529
Random Forest	TF-IDF	0.9533	0.9529	0.9533	0.9531
XGBoost	TF-IDF	0.9566	0.9563	0.9566	0.9564
Logistic Regression	Bag Of Words	0.9527	0.9523	0.9527	0.9525
Naïve Bayes	Bag Of Words	0.9319	0.9457	0.9319	0.9351
SVM	Bag Of Words	0.9581	0.9578	0.9581	0.9579
Random Forest	Bag Of Words	0.9569	0.9566	0.9569	0.9567
XGBoost	Bag Of Words	0.9584	0.9585	0.9584	0.9584
Logistic Regression	Word2Vec	0.9212	0.9194	0.9212	0.9200
SVM	Word2Vec	0.9292	0.9311	0.9292	0.9300
Random Forest	Word2Vec	0.9367	0.9369	0.9367	0.9368
XGBoost	Word2Vec	0.9393	0.9397	0.9393	0.9395

These results offer a useful opportunity to investigate how well various algorithms perform when classifying tweets of emergencies. Strong accuracy is demonstrated by machine learning models like XGBoost and SVM, which achieve above 95% accuracy when the Bag of Words (BoW) and TF-IDF approaches are used. This implies that word frequency-based techniques are useful for identifying emergency messages, most likely due to the fact that urgent tweets frequently follow particular patterns. Word2Vec-based models' comparatively worse performance of the Word2Vec-based models raises the possibility that pre-trained word embeddings could not adequately convey the context or urgency of these signals. Although SVM and XGBoost perform well, they may have trouble understanding more complex or context-dependent emergency tweets because of their emphasis on basic frequency patterns.

On the other hand, the deep learning models—BLSTMA with FastText embeddings, in particular—performed better, achieving 96.64% accuracy. This implies that FastText embeddings help the model understand context, allowing it to identify crises even in various language usages. Deep learning models also have stronger recall, which lowers the likelihood of overlooking important signals. Deep learning models exhibit a remarkable ability to navigate the intricate and informal nature of Twitter discourse, which improves their dependability for real-world emergency detection, even though machine learning models are more efficient in terms of processing speed and efficiency.

Conclusion

The results showed that while the BLSTMA model had the best accuracy (96.64%) and F1 score (0.9116), conventional machine learning techniques like XGBoost and SVM. Using Bag of Words vectorisation, SVM obtained 95.81% accuracy and an F1 score of 0.9579, whereas XGBoost earned 95.84% accuracy and an F1 score of 0.9584. By demonstrating the usefulness of the BLSTMA model in real-time disaster response and the complementary advantages of conventional approaches in the analysis of complex disaster data, these findings highlight the significance of customising machine learning and deep learning approaches to particular tasks.

Even though deep learning models demonstrated excellent generalisation skills, the need for further improvement is highlighted by their comparatively poorer accuracy and recall compared to conventional techniques. Additionally, the study's conclusions are not as broadly applicable to other languages or catastrophe situations due to its dependence on a dataset of Turkish tweets. To improve these models' worldwide applicability, future studies should examine how well they can adapt to multilingual datasets and various settings.

Furthermore, there are difficulties in processing large data streams (e.g., the vast number of messages shared on social media during disaster events), noisy data (e.g., misleading, incomplete, or irrelevant information within the shared messages), and the requirement for quick categorisation (e.g., the need to analyse and classify data in real time to support immediate decision-making) when implementing these models in real-time disaster response systems. Overcoming these obstacles requires combining real-time processing capabilities (e.g., enabling models to process and classify incoming data as it is collected) and increasing computing efficiency (e.g., optimising the speed and resource usage of models for faster performance). By addressing these issues, it will be possible to create frameworks for evaluating social media data connected to disasters that are more reliable and scalable, greatly improving emergency response and disaster management initiatives throughout the globe.

Ethics Committee Approval: Ethics approval was not required, as the analysis relied solely on publicly available Twitter data, ensuring anonymity and no use of personal information.

Peer-review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- M.S.; Data Acquisition- M.S.; Data Analysis/Interpretation- M.S., A.Ö.; Drafting Manuscript- M.S., A.Ö.; Critical Revision of Manuscript- M.S., A.Ö.; Final Approval and Accountability- M.S., A.Ö.; Technical or Material Support- M.S., A.Ö.; Supervision- M.S., A.Ö.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: The authors declared that this study has received no financial support.

References

Al Jazeera (2023). "Quake victims stuck under rubble take to social media for help." Available: https://www.aljazeera.com/news/2023/2/7/quake-victims-stuck-under-rubble-take-to-social-media-for-help

Euronews (2023). "How Twitter helped find survivors trapped beneath rubble after Turkey's earthquakes." Available: https://www.euronews.com/next/2023/02/10/how-twitter-helped-find-survivors-trapped-beneath-rubble-after-turkeys-earthquakes

Powers, C. J., Devaraj, A., Ashqeen, K., Dontula, A., Joshi, A., Shenoy, J., & Murthy, D. (2023). Using artificial intelligence to identify emergency messages on social media during a natural disaster: A deep learning approach. *International Journal of Information Management Data Insights*, *3*(1), 100164. https://doi.org/10.1016/j.jjimei.2023.100164

Bhere, P., Upadhyay, A., Chaudhari, K., & Ghorpade, T. (2020). Classifying Informatory Tweets during Disaster Using Deep Learning. *ITM Web of Conferences*, *32*, 03025. https://doi.org/10.1051/itmconf/20203203025

Kumar, A., Singh, J. P., & Saumya, S. (2019). A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification. In *2019 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 1-6). IEEE. https://doi.org/10.1109/R10-HTC47129.2019.9042443

Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. S. (2021). Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises. *International Journal of Disaster Risk Reduction, 55*, 102101. https://doi.org/10.1016/j.ijdrr.2021.102101

Madichetty, S., & Sridevi, M. (2020). A stacked convolutional neural network for detecting the resource tweets during a disaster. *Multimedia Tools and Applications, 80*, 3927-3949. https://doi.org/10.1007/s11042-020-09873-8

Sit, M.A., Koylu, C., & Demir, I. (2019). Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma. *International Journal of Digital Earth, 12*, 1205-1229. https://doi.org/10.1080/17538947.2018.1563219

Mijwel, M. M. (2015, April). *History of Artificial Intelligence*. *Computer Science*, 3–4

Kang, Y., Cai, Z., Tan, C., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. (pp. 1–20)

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408. https://doi.org/10.1037/h0042519

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. (pp. 168–171).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. (pp. 373–383).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). IEEE. https://doi.org/10.1109/ICASSP.2013.6638947

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 207–212). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2034

Zhong, G., Lin, X., Chen, K., Li, Q., & Huang, K. (2020). Long short-term attention. In *Lecture Notes in Computer Science* (pp. 45–54). Springer. https://doi.org/10.1007/978-3-030-39431-8_5

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” in *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, Nov. 1997, doi: 10.1109/78.650093

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*.