electronics

MDPI

Article

Neural Network-Based Sentiment Analysis and Anomaly Detection in Crisis-Related Tweets

Josip Katalinić and Ivan Dunđer *

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 10000 Zagreb, Croatia
* Correspondence: idundjer@ffzg.unizg.hr

Abstract: During crises, people use X to share real-time updates. These posts reveal public sentiment and evolving emergency situations. However, the changing sentiment in tweets coupled with anomalous patterns may indicate significant events, misinformation or emerging hazards that require timely detection. By using a neural network, and employing deep learning techniques for crisis observation, this study proposes a pipeline for sentiment analysis and anomaly detection in crisis-related tweets. The authors used pre-trained BERT to classify tweet sentiment. For sentiment anomaly detection, autoencoders and recurrent neural networks (RNNs) with an attention mechanism were applied to capture sequential relationships and identify irregular sentiment patterns that deviate from standard crisis talk. Experimental results show that neural networks are more accurate than traditional machine learning methods for both sentiment categorization and anomaly detection tasks, with higher precision and recall for identifying sentiment shifts in the public. This study indicates that neural networks can be used for crisis management and the early detection of significant sentiment anomalies. This could be beneficial to emergency responders and policymakers and support data-driven decisions.

Keywords: sentiment analysis; anomaly detection; neural networks; deep learning; crisis monitoring; X platform; social media analytics

Academic Editors: Patrick Siarry and Junaid Rashid

Received: 3 April 2025
Revised: 26 May 2025
Accepted: 29 May 2025
Published: 2 June 2025

Citation: Katalinić, J.; Dunđer, I. Neural Network-Based Sentiment Analysis and Anomaly Detection in Crisis-Related Tweets. *Electronics* **2025**, *14*, 2273. https://doi.org/10.3390/electronics14112273

1. Introduction

Social media platforms like X (formerly known as Twitter) are widely used to broadcast information and organize support during disasters [1]. During disasters, Twitter is used for delivering time-sensitive information that facilitates situational awareness of the event [2]. Understanding public sentiment in these crisis-related tweets (i.e., messages posted on X) can provide valuable insights for emergency responders and decision-makers [3]. Analysis of social media sentiment provides real-time information on public opinion and emotional reactions during disasters. This aspect of real-time analysis has gained importance in crisis management [4]. Understanding typical sentiment patterns provides the context necessary to detect anomalies. Kumar et al. [5] stated that an anomaly is an observation that deviates so much from other observations that it raises the suspicion that it was generated by a different mechanism. Anomaly detection techniques such as logistic regression, naïve Bayes, and k-nearest neighbors (KNN) have been extensively studied [6,7], but the unique high-volume and fast-paced nature of social media has attracted increased interest in exploring new anomaly detection methods tailored to social media platforms and data [8]. For crisis management, timely anomaly detection can provide information crucial to support decision-making. Analysts and decision-makers can leverage such data to obtain an assessment of the scope of the events and effectively allocate resources [9].

Traditional machine learning (ML) methods often require manual feature engineering, which can be time-consuming and error-prone [8]. To address these challenges, deep neural networks are used to detect anomalies within data, which can provide deeper insights into the context of anomalies [10]. Neural network-based transformer models like BERT [11] have removed manual intervention, and achieved high accuracy in the field of natural language processing (NLP) by capturing context and nuances in messages [12]. Likewise, different neural network architectures such as recurrent networks [13] and autoencoders [14] can model complex temporal patterns for anomaly detection, potentially spotting subtle deviations.

This study proposes a neural network-based framework that integrates sentiment analysis and anomaly detection for crisis-related tweets. This is achieved by leveraging BERT for sentiment classification, and using sequential autoencoders with an attention mechanism for unsupervised anomaly detection. The goal is to automatically classify the sentiment of crisis tweets (e.g., positive, negative, neutral), and simultaneously identify anomalous (irregular) spikes or deviations (outliers) in the Twitter stream that could signal critical events. By combining these two capabilities, emergency decision-makers could gain a holistic overview by monitoring public sentiment trends, while also being alerted to unusual events (social media content) in real time. The remainder of this paper is structured as follows: related work is discussed in Section 2, the methodology is presented in Section 3, research results are provided in Section 4, implications are stated in Section 5, and a conclusion with future directions is provided in Section 6.

2. Related Work

Sentiment analysis on social media has been widely researched [15–17], including the sentiment analysis of natural disasters [18]. The ability to accurately extract sentiment and detect anomalies is essential to understanding and improving crisis response. Other dimensions of crisis informatics have been explored in recent works, such as integrating image analysis with text for disaster assessment [19], and leveraging geospatial information from tweets [20].

2.1. Sentiment Analysis in Crisis-Related Tweets

Early works applied classical machine learning algorithms to Twitter data. Suhasini and Srinivasu [21] detected sentiment in tweets using supervised learning, comparing k-nearest neighbors (KNN) and naïve Bayes classifiers. Jayakody and Kumara [22] analyzed product review tweets using support vector machines (SVM), logistic regression, and KNN with standard text features. Recent research has moved towards neural network models for tweet sentiment analysis [23]. Convolutional neural networks (CNN) and recurrent neural networks (RNN) have been employed to capture context from tweet text. Stojanovski et al. [24] proposed a CNN-based feature extractor combined with various classifiers for emotion identification in tweets. Devlin et al. [11] stated that the emergence of language models such as BERT (Bidirectional Encoder Representations from Transformer) has improved the reliability of NLP tasks, including sentiment classification. BERT's bidirectional transformer architecture enables the entire input text sequence to be read simultaneously, and context captured in both directions, unlike earlier left-to-right or right-to-left models. Myint et al. [25] applied BERT and its variants to a crisis-related dataset with similar results in both "negative" and "neutral" sentiment categories, achieving F1 scores of 0.92 and 0.82, respectively. In the "positive" sentiment category, BERT and BERTweet [26] performed similarly, with an F1 score of 0.91, while RoBERTa (Robustly Optimized BERT Pretraining Approach) [27] followed closely with an F1 score of 0.90. A recent hybrid model by Kanungo and Jain [4] combined gated recurrent units (GRUs) [28]

for selectively retaining relevant information, and long short-term memory (LSTM) [29] to capture dependencies in sequential data into a so-called "G-LSTM" network for sentiment analysis of disaster tweets, reporting over 90% accuracy on a custom dataset.

2.2. Anomaly Detection in Crisis-Related Tweets

Anomaly detection on X and other social media has been studied from the point of view of event detection. According to Peng et al. [30], the event detection task is more challenging than traditional text mining or social media mining, since a general social event is a meaningful and influential combination of social messages in the open domain. Events often contain event-related heterogeneous elements, such as location, person, organization, relationships, date and time, keywords, etc. In particular, social message content is always overlapping, where the noisy nature of the message stream makes traditional outlier detection technologies unsuitable for the semantically rich task of event detection. Patel et al. [31] introduced real-time sentiment-based anomaly detection, where incoming tweets were first classified by sentiment (positive, neutral, negative), and then each sentiment stream was monitored for anomalous (abnormal) surges. Their RSAD (Real-Time Sentiment-Based Anomaly Detection) system detects spikes in the number of tweets that deviate from the baseline, using techniques such as exponentially weighted moving averages. They underscore the value of incorporating sentiment signals into event detection—surges in highly negative tweets, for example, may indicate a worsening situation or public outrage that demands attention. Other research has studied content anomalies such as messages that are semantically unusual compared to typical chatter. For instance, Kumar et al. [5] developed an anomaly detection framework for Twitter (now known as X) that involved topic modeling and clustering. Their workflow used latent Dirichlet allocation (LDA) to identify topics in tweets, sentence-transformer embedding and k-means clustering to cluster tweets, thus flagging an unusual cluster as a potential anomaly event.

Deep learning (DL) has also been applied to detect anomalies in social media. Some works use autoencoders to learn a representation of normal social media behavior, and detect outliers via reconstruction error [32]. Others use sequence models such as LSTMs [33] to predict future data points (e.g., number of tweets or engagement metrics), and signal anomalies when predictions differ significantly. Researchers have enhanced such models with attention mechanisms to improve focus on relevant features or time steps. For instance, an attention-based LSTM autoencoder was proposed by Do et al. [34] for time-series anomaly detection, allowing the model to capture important parts of the sequence during reconstruction.

In the context of crisis events and crisis informatics, anomaly detection remains challenging due to the concept of drift—what is normal can change rapidly as an event occurs. These studies show that combining textual analysis with anomaly detection can improve the detection of anomalies over purely statistical methods. Building on these ideas, this research uses a neural autoencoder (with an attention mechanism) directly on tweet content sequences to detect anomalous patterns, alongside a sentiment classifier. This dual approach aims to simultaneously measure public sentiment and capture unusual signals in the data stream.

3. Materials and Methodology

For the purpose of this study, the authors used the Turkey and Syria Earthquake Tweets dataset, which is available on Kaggle [35]. The dataset contains 472,399 tweets relating to the earthquake that struck Turkey and Syria on 6 February 2023, and ends with tweets on 21 February 2023. The dataset captures real-time, user-generated tweets reflecting interactions, public responses, and reactions during the event. All tweets included a language

metadata field. The dataset was filtered to include only English-language tweets, yielding 189,626 tweets for this analysis (the filtering was conducted via the provided language label). Location information was excluded from all subsequent modelling because 96% of the tweets lacked geolocation. It should be noted that this dataset did not come with sentiment annotations. Therefore, sentiment labels were assigned using a pre-trained BERT sentiment model, rather than manual human annotations.

3.1. Data Preprocessing

Data preprocessing is an important step to increase the quality and usability of tweet data. The data preprocessing workflow is visualized in Figure 1. All the following steps were accomplished with Python version 3.8.10:

1. Filtering English language: by using pandas, a Python data analysis library, in the language column, only rows where the language field had the value "en" were filtered. This step was necessary to increase the reliability of the pre-trained BERT model for sentiment analysis [36]. After this filtering, 189,626 tweets out of 472,399 tweets were filtered as English text.
2. Text lowercasing: all tweets were converted to lowercase; according to Hickman et al. [37], lowercasing tends to be beneficial because it reduces data dimensionality, thereby increasing statistical power, and usually does not reduce validity.
3. Stop word removal: common English (function) words such as "and", "is", "I", "am", "what", "of", etc. were removed by using the Natural Language Toolkit (NLTK). Stop word removal has the advantages of reducing the size of the stored dataset and improving the overall efficiency and effectiveness of the analysis [38].
4. URLs removal: all URLs were removed from tweets, since the text of URL strings does not necessarily convey any relevant information, and can therefore be removed [39].
5. Duplicate removal: all duplicate tweets were removed to eliminate redundancy and possible skewing of the results.

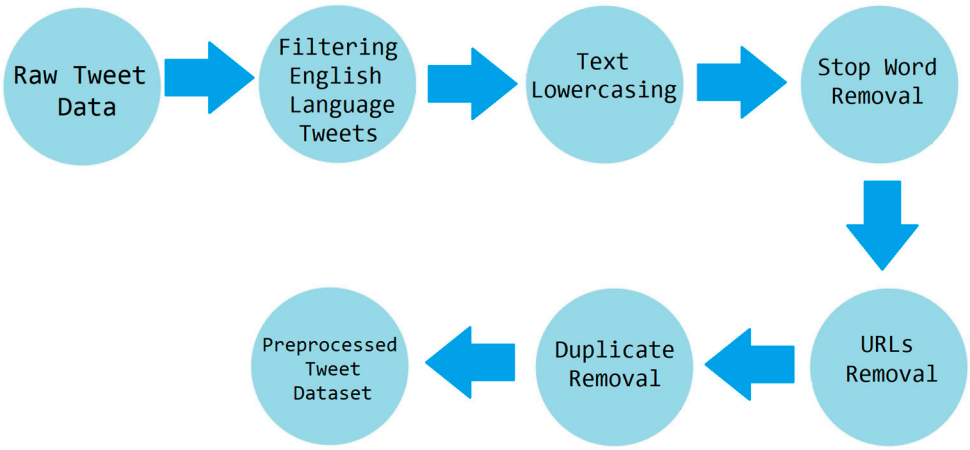

Figure 1. Tweet data preprocessing workflow.

3.2. Methods

After data preprocessing, two different analyses (sentiment analysis and anomaly detection) were carried out using five different methods (Figure 2). The first analysis (sentiment analysis) was carried out using a neural network with the pre-trained transformer-based BERT model. A traditional machine learning approach with logistic regression was used for the same task to contrast the neural network for sentiment analysis. The second analysis was anomaly detection, where the neural network was divided into two categories, autoencoder neural network and LSTM neural network, with an integrated

attention mechanism. For the traditional machine learning approach, an isolation forest was used for anomaly detection.

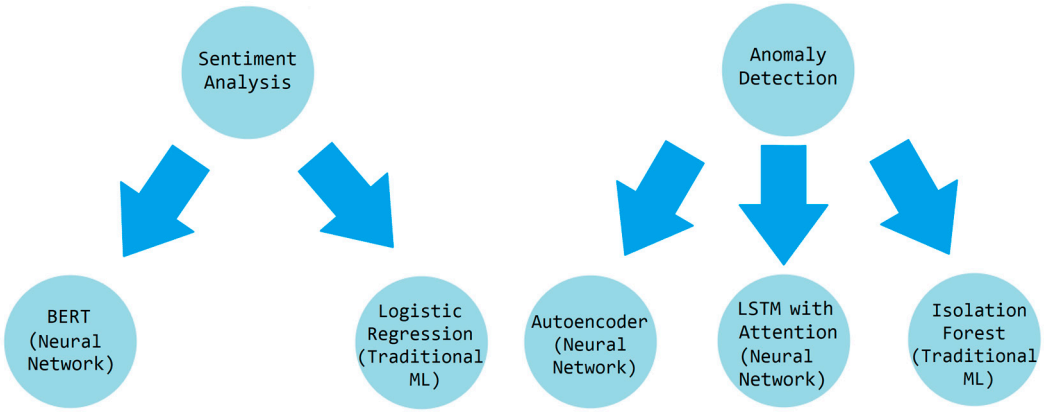

Figure 2. Workflow of tweet data preprocessing.

Sentiment analysis was performed utilizing a pre-trained transformer-based BERT model, specifically the "nlptown/bert-base-multilingual-uncased-sentiment". This model is a fine-tuned version of "bert-base-multilingual-uncased", which is optimized for sentiment analysis across six languages: English, Dutch, German, French, Spanish and Italian [40]. Tweets were tokenized using the AutoTokenizer from HuggingFace Transformers, truncated to a maximum length of 512 tokens [41]. The model predicted sentiment scores across five classes representing very negative to very positive sentiments. These categorical outputs were then converted to a continuous polarity scale ranging from −1 (strongly negative) to +1 (strongly positive) to facilitate the temporal analysis of sentiment fluctuations.

An autoencoder neural network was designed and trained to detect anomalies based on deviations in tweet sentiment patterns. The input data was structured into sequences of polarity scores. The autoencoder was implemented as a fully connected feedforward network with a three-layer encoder and symmetric decoder. The encoder consisted of a hidden layer with 64 neurons followed by a 16-neuron bottleneck, using rectified linear unit (ReLU) activations for encoding and decoding [42]. Reconstruction errors (mean squared error between actual and reconstructed sequences) were calculated, and tweets with errors above the 95th percentile threshold were flagged as anomalies. An LSTM neural network with an integrated attention mechanism was implemented to detect anomalies based on prediction errors. Input sequences of polarity scores were processed through LSTM layers, and attention layers were applied to selectively weigh temporal dependencies within the sequences. The LSTM with attention included a single-layer LSTM model with a hidden size of 32, followed by an attention mechanism. Both models were trained for 10 epochs using the Adam optimizer (learning rate was set to 0.001), with a batch size of 32 and mean squared error (MSE) loss. Sentiment polarity scores were normalized using MinMax scaling to the [0,1] range. The model's output was a prediction of subsequent sentiment scores. Anomalies were identified when prediction errors exceeded a threshold set at the 95th percentile, highlighting sudden or extreme shifts (changes) in sentiment.

Logistic regression served as a traditional machine learning approach for sentiment classification in this research. The model used L2 regularization with the default inverse regularization strength (C = 1.0), and was trained for 200 iterations. Term frequency-inverse document frequency (TF-IDF) vectorization was applied to convert textual tweet data into numeric vectors. A logistic regression model was then trained on these vectors to classify sentiment into binary classes (positive or negative sentiment). Model training included standard parameter tuning and cross-validation to optimize model performance. For

anomaly detection, the authors utilized the isolation forest algorithm with 100 estimators and a contamination rate of 0.05, aligning with the 95th percentile threshold used in this study's neural network models. The model was applied directly to polarity scores, identifying anomalous sentiment scores by isolating data points based on their unique characteristics relative to the majority. Anomalies were defined as data points that required fewer splits to isolate in a binary search tree structure, allowing for efficient detection of sentiment outliers. All experiments were executed on a workstation equipped with an RTX 4090 GPU, an Intel i9-14900K CPU, 64 GB RAM and Windows 11.

4. Results

The data analysis results are discussed in two task-related subsections: sentiment analysis and anomaly detection. These tasks provided insight into public sentiment, and identified anomalous sentiment patterns corresponding to significant real-time events within the crisis scenario. The analysis was based on a dataset of 189,625 English-language tweets, ensuring comprehensive coverage of user-generated crisis-event-related content. The results demonstrate the ability of advanced sentiment analysis and anomaly detection techniques to uncover meaningful patterns and deviations in social media discourse during crises.

4.1. Sentiment Analysis

Sentiment analysis was applied to a total of 189,626 English-language tweets with polarity scores ranging from -1 (most negative) to +1 (most positive). The tweets captured diverse emotional responses and reactions, including varying levels of public concern, panic, support and empathy during the crisis. The overall polarity indicates a generally negative sentiment of -0.28. Despite the dominance of negative sentiments in crisis-related tweets, expressions of positivity were evident, reflecting community solidarity and supportive engagements alongside signals of distress.

A visualization of daily average sentiment polarity scores (Figure 3) reveals distinct temporal variations, illustrating fluctuations within predominantly negative sentiment. Sentiment scores showed clear peaks and fluctuations, reflecting public reactions as the crisis unfolded in real time. Periods exhibiting sharp declines suggest intensified negative emotional responses, perhaps triggered by escalating events or worsening news, while periods of relatively less negative sentiment may correspond to phases of crisis stabilization or expressions of community solidarity and relief.

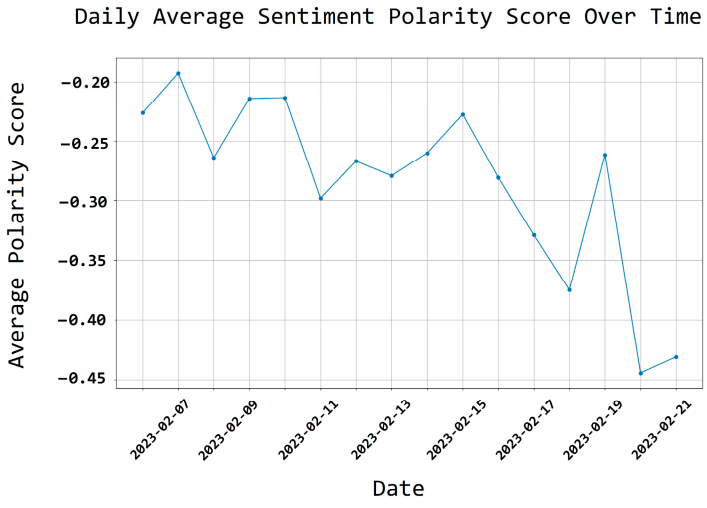

Figure 3. Daily average sentiment polarity score over time.

4.2. Anomaly Detection

Anomaly detection was performed using two different methodologies: an autoencoder and an LSTM model enhanced by an attention mechanism. To identify anomalies, both methods employed a threshold set at the 95th percentile of reconstruction errors (autoencoder) or prediction errors (LSTM). Since anomalies are assumed to be rare [43] the 95th percentile was used as a threshold for unsupervised anomaly detection, with the assumption that 5% of the data correspond to anomalies. According to Cai et al. [44], the reconstruction error in anomaly detection using an autoencoder serves as the anomaly score, indicating the extent of deviation from normality. Large reconstruction errors usually indicate anomalies as the model struggles to accurately reconstruct atypical patterns. The prediction error in the context of LSTM, according to Wang et al. [45], corresponds to the deviation between the actual and predicted values. High prediction errors imply that the observed data deviate significantly from the expected patterns indicating an anomaly.

Setting the anomaly threshold at the 95th percentile of reconstruction or prediction errors resulted in 9481 anomalies per method, and visualizing such a large number (18,962 anomalies combined for both methods) would significantly reduce the readability, interpretability and practical usefulness of the presented graphs. Therefore, to illustrate and analyze the anomaly patterns, specifically the subset of anomalies, 80 anomalies from each detection method (autoencoder and LSTM) were randomly selected (Figure 4). The autoencoder-based method detected anomalies by analyzing reconstruction errors from expected sentiment patterns. Tweets that exceeded the threshold typically corresponded to significant events or breaking news, such as early reports of large casualty numbers or emotional descriptions of disasters. The LSTM model with attention complemented the autoencoder by highlighting anomalies based on prediction errors, effectively detecting rapid shifts in sentiment. The LSTM model was particularly sensitive to emotionally charged content, successfully flagging tweets with heightened emotional content such as impactful images or distress signals.

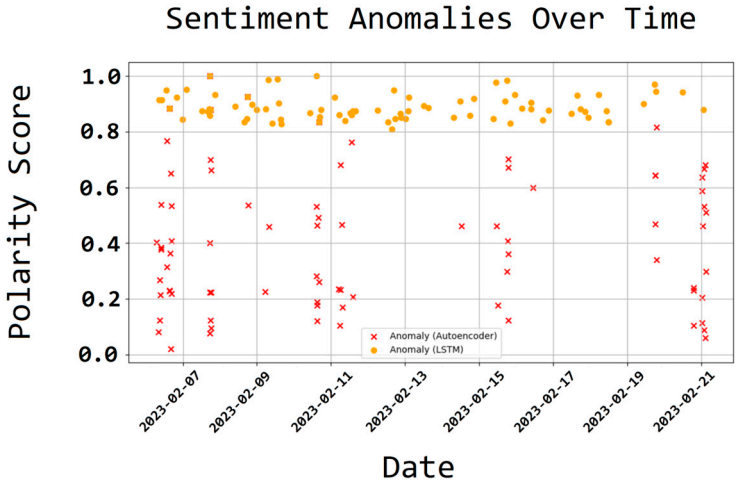

Figure 4. Sentiment anomalies over time.

In Figure 4, it can be seen that the anomalies detected by the autoencoder (red crosses) are mainly located in the lower to middle part of the graph, reflecting negative to slightly positive sentiments. Given that the autoencoder recognizes anomalies on the reconstruction error, it has detected sentiments that deviated from the typical neutral or negative sentiments characteristic of crisis-related tweets. In contrast, the LSTM anomalies (yellow circles) are located in the upper part of the graph, reflecting high polarity anomalies. The LSTM model recognized anomalies in the higher polarity range, indicating unusually strong positive sentiments that deviated from the predicted sentiment trajectories.

To address model interpretability, the authors applied LIME (local interpretable model-agnostic explanations) to explain the anomaly scores generated by both the autoencoder and LSTM. A representative sequence was selected based on its reconstruction error being closest to the 95th percentile anomaly threshold. In the LIME plots (Figures 5 and 6), each feature is labeled as t-i, where i denotes the number of time steps before the current tweet (t-0). Bars indicate how much each sentiment input influenced the anomaly score, with green representing features that increased anomaly likelihood, and red representing those that reduced it.

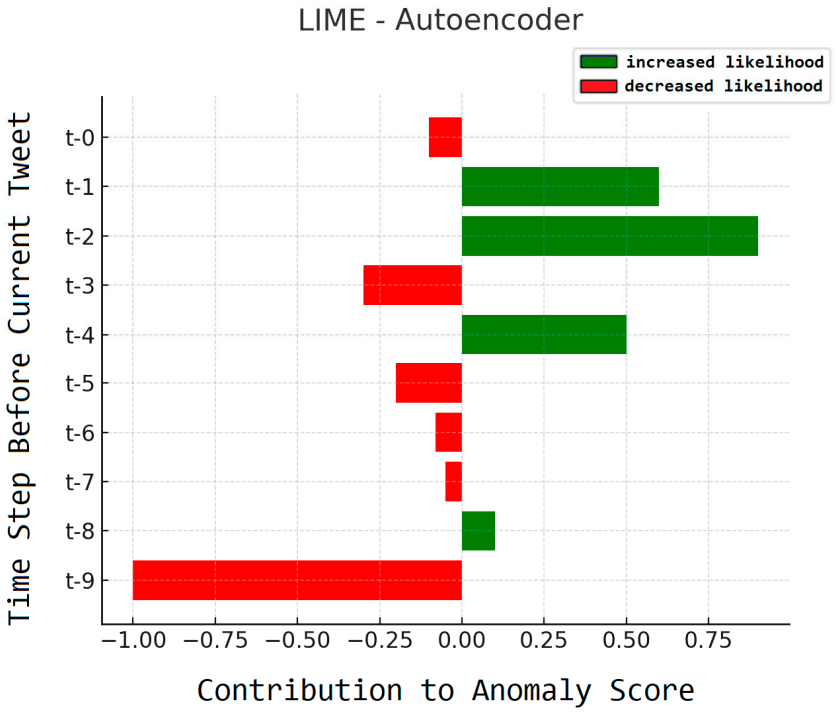

Figure 5. LIME—autoencoder.

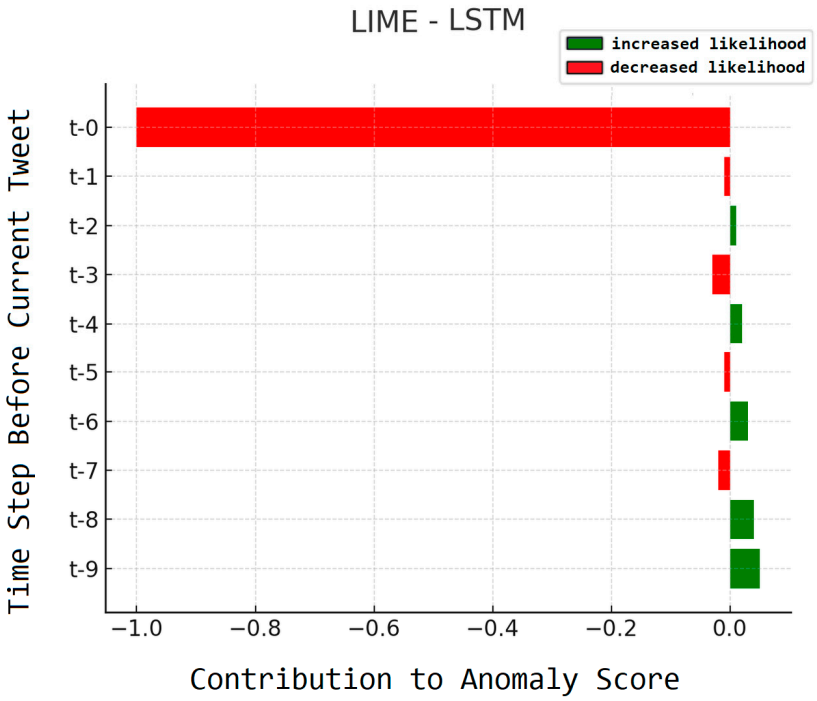

Figure 6. LIME—LSTM.

The autoencoder emphasized longer-term sentiment history, with t-2 and t-1 (recent high sentiment) increasing the anomaly score, while older inputs like t-9 reduced it. In contrast, the LSTM with attention focused on short-term dynamics: t-0 (the most recent sentiment) had the strongest negative contribution, signaling normal behavior, and reducing anomaly likelihood. This interpretability is especially valuable for applications involving policymakers and emergency responders, where understanding the basis of a flagged anomaly is crucial for trust and actionable response.

4.3. Comparative and Temporal Analysis of Anomalies

A comparative analysis of anomalies showed that anomalous tweets had a higher average polarity compared to normal tweets. For anomalous tweets, a polarity of 0.639 was detected, while normal tweets had an average polarity of 0.364. The distribution of polarity scores (Figure 7) illustrates significant differences between anomalous and normal tweets. The blue bars correspond to the distribution of polarity scores for normal tweets, and the red bars correspond to the distribution of polarity for anomalous tweets. The lines illustrate the overall trend in the distribution of polarity scores for each group. Anomalous tweets show polarity scores with extremes on the scale, displaying strongly positive or significantly negative sentiments. This distribution is in clear contrast to the more neutral distribution of polarity scores for normal tweets. Such a distribution demonstrates the sensitivity and effectiveness of anomaly detection methods by identifying emotionally charged tweets.

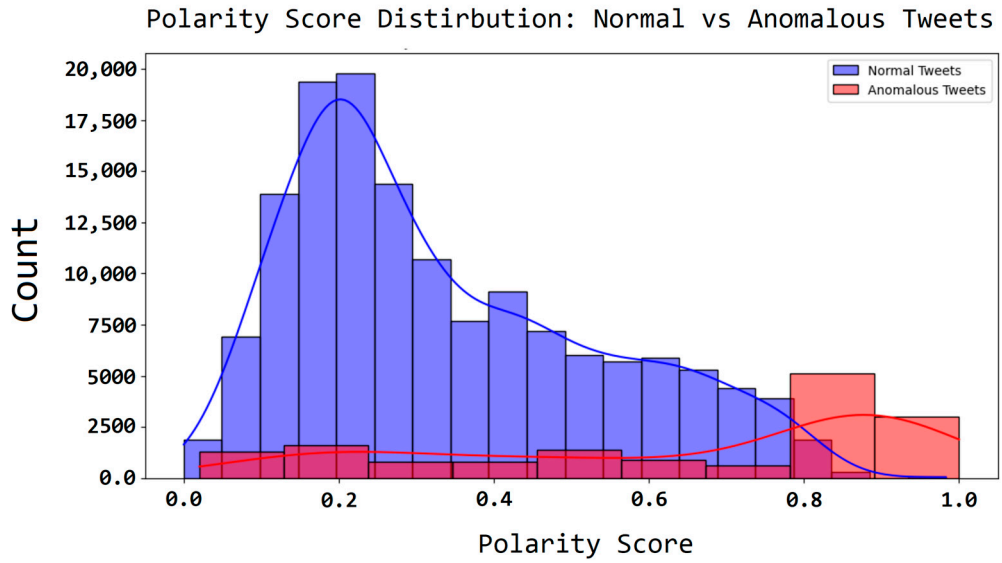

Figure 7. Polarity score distribution: normal vs. anomalous tweets.

The temporal analysis of anomalous tweets provides additional context by highlighting their alignment with real-time crisis developments. Visualizations comparing daily average polarity scores between normal and anomalous tweets (Figure 8) show that anomalous tweets coincided with significant real-world crisis events. Such anomalous tweets reflected shifts in sentiment dynamics during critical periods. Sharp peaks and troughs in anomalous sentiment were closely aligned with key events such as initial news reports of strong earthquake impacts, public calls for help or expressions of solidarity and support.

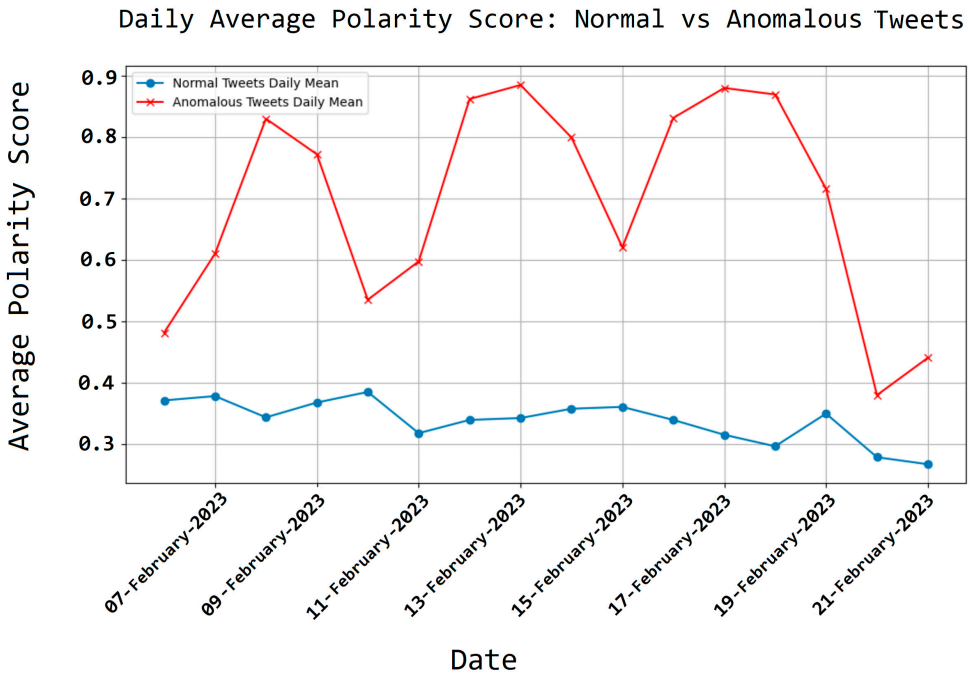

Figure 8. Daily average polarity score: normal vs. anomalous tweets.

To better understand the interpretation and context of individual anomalous tweets, a qualitative review was applied. By randomly selecting anomalous tweets, the analysis confirmed extremely emotional or contextually intensive content. Such tweets included examples like "⚠⚠⚠⚠ #earthquake #IAMUNDERTHEDEBRIS", which show urgent signals for help, while emotionally supportive tweets like "solidarity and unity" indicate strong collective empathy during the crisis. When observing surrounding tweets, they had neutral sentiments, emphasizing the clear deviation of identified anomalies from their contextual baselines.

4.4. Comparing Neural Network with Traditional ML Approaches

To evaluate the performance of neural network models compared to traditional machine learning methods, a random subset of 2000 tweets was manually labeled and used as a baseline. Each tweet sentiment polarity was manually labeled by three independent annotators. The final polarity score was determined by averaging these three scores. Another 2000 random tweets were selected to assess anomaly detection. Here, 500 out of 1500 were randomly selected anomalies, and they were evaluated by the same three independent annotators. This targeted sampling was introduced because anomalies in the dataset were rare, potentially leading evaluators to the wrong conclusion due to expectation bias. A majority voting approach was introduced where the label "anomalous" or "normal" was chosen by at least two annotators.

For sentiment analysis, a comparison was made between the previously used BERT neural network model and logistic regression. According to Premasudha and Rampalli [46], logistic regression is a statistical model that, despite its name, is commonly used for classification tasks, especially for binary classification. Logistic regression can estimate the probabilities of a particular piece of text being positive or negative by using a logistic function. Jahan et al. [47] state that the model's advantages lie in its simplicity, interpretability and efficiency in high-dimensional spaces, making it suitable for text classification tasks. However, it assumes a linear relationship between features, which may not always be true in complex sentiment expressions.

The neural network-based sentiment analysis performed better (Table 1) compared to traditional machine learning approaches (Table 2) for both classes used in sentiment analysis (positive and negative). The sentiment analysis based on the BERT neural network showed a high score for sentiment classification with a precision of 91%, a recall of 85% and an F1 score of 88% for positive sentiments. These results outperformed logistic regression, which achieved an F1 score of 54%. In comparison, logistic regression showed a low score for sentiment classification, especially in identifying positive sentiments, and this was particularly evident with a low recall of 41%.

Table 1. Neural Network (BERT) sentiment classification.

Class	Precision	Recall	F1 Score
0 (negative)	0.88	0.9	0.89
1 (positive)	0.91	0.85	0.88

Table 2. Traditional ML (logistic regression) sentiment classification.

Class	Precision	Recall	F1 Score
0 (negative)	0.75	0.82	0.78
1 (positive)	0.78	0.41	0.54

In the anomaly detection task, the autoencoder (Table 3) and the LSTM neural network method with attention (Table 4) outperformed traditional machine learning approaches (Table 5), similarly to the sentiment classification task. Anomaly detection based on the autoencoder neural network achieved a high accuracy of 89%, with solid performance in identifying normal sentiment patterns (86% recall). Despite a mild recall for anomalies of 67%, its high precision of 88% signifies reliable identification of true anomalies. The LSTM neural network with attention achieved lower results in all classes compared to the autoencoder, with a precision of 83% for identifying normal sentiment patterns and 82% for anomalies. The traditional isolation forest approach compared to neural network methods demonstrated significantly lower precision (69%) and similar recall (64%), making it less reliable for anomaly detection with a higher risk of false positives.

Table 3. Neural network (autoencoder) anomaly detection.

Class	Precision	Recall	F1 Score
0 (negative)	0.89	0.86	0.87
1 (positive)	0.88	0.67	0.76

Table 4. Neural network (LSTM with attention) anomaly detection.

Class	Precision	Recall	F1 Score
0 (negative)	0.83	0.87	0.85
1 (positive)	0.82	0.65	0.73

Table 5. Traditional ML (isolation forest) anomaly detection.

Class	Precision	Recall	F1 Score
0 (negative)	0.81	0.83	0.82
1 (positive)	0.69	0.64	0.66

To assess the relative impact of key components, the authors performed controlled ablations targeting BERT fine-tuning, autoencoder depth, and LSTM attention layers. Five paired train/validation splits were evaluated for each comparison, and statistical significance was assessed with two-tailed paired *t*-tests ($\alpha = 0.05$). Results are summarized in Table 6. Only the inclusion of the attention mechanism yielded a statistically significant improvement (t = 4.57, $p = 0.011$). Neither BERT fine-tuning ($p = 0.173$) nor additional autoencoder depth ($p = 0.385$) produced significant gains.

Table 6. Ablation effects on the model's performance.

Component	Variant Full	Variant Ablated	t	*p*
BERT encoder	fine-tuned	frozen encoder	1.66	0.173
Autoencoder depth	3-layers deep	1-layer deep	0.98	0.385
LSTM attention	LSTM + attention	LSTM	4.57	0.011

5. Discussion

Experimental findings have demonstrated better results when using neural network-based methods (BERT and autoencoder) for sentiment analysis and anomaly detection in crisis-related tweets compared to traditional machine learning models such as logistic regression and isolation forest. Sentiment classification results showed substantial improvements in precision and recall when utilizing BERT compared to logistic regression. These improvements can be attributed to BERT's deep transformer architecture, which effectively captures context, semantic nuances and complex linguistic structures within tweet data [48]. The authors specifically chose BERT due to its proven effectiveness and the availability of pretrained weights for sentiment classification. Alternative transformer models (such as RoBERTa or domain-specific models like BERTweet) might further enhance sentiment accuracy [26,27]. In addition, evaluating these alternative models on the dataset was outside the scope of the current study, but the authors mention it as a potential improvement. Logistic regression, using simple TF-IDF vectorization, struggled to accurately classify positive sentiment, resulting in notably lower recall and precision. The results of these metrics indicate greater robustness and reliability of the neural network method in accurately capturing the emotional dynamics reflected in tweets.

In anomaly detection tasks, neural network methods (autoencoder and LSTM with an attention mechanism) provided better performance compared to traditional machine learning methods. The authors opted for an autoencoder-based anomaly detector because it does not require labeled anomalies, and because it can automatically learn complex patterns of "normal" sentiment. In contrast, standard time-series outlier detection techniques (e.g., control charts or forecasting models) often require manual feature engineering or assumptions of stationarity that do not hold in rapidly changing crisis data [49,50]. In particular, the autoencoder and LSTM-based anomaly detection models had higher precision and moderate recall for anomalous tweets, effectively identifying tweets with emotionally charged content or critical crisis-related information updates. The traditional isolation forest approach, although capable of recognizing anomalies to a certain extent, demonstrated lower precision and recall, leading to higher false-positive rates with limited practical effectiveness. Given the critical nature of crises, the precision of anomaly detection is especially vital, reinforcing the advantages of neural networks in practical crisis management.

Experimental analyses also revealed significant polarity score differences between anomalous and normal tweets. These differences highlighted that the anomalies included either strongly positive sentiments reflecting supportive community responses or strongly

negative tweets reporting urgent distress. The temporal analysis demonstrated that the anomalies detected through neural networks closely corresponded to the development of real-world crisis situations, such as urgent calls for help from involved victims or emotionally supportive tweets indicating strong collective empathy during the crisis. In terms of practical implications, these findings underscore the value of employing neural network-based methods for real-time crisis management. Such methods allow for the early detection of sentiment shifts, potentially enabling rapid responses from emergency services, policymakers, and humanitarian organizations. However, the moderate recall observed in anomaly detection indicates potential for further improvement. Integrating multimodal data [51] or exploring ensemble methods [52] could improve sensitivity and recall without sacrificing precision.

A primary limitation of this study was its reliance solely on textual data within tweets, overlooking other potentially valuable multimodal data, such as images, videos or user metadata. These data points could provide richer context for anomaly detection. It should be noted that this current evaluation was conducted on historical data, and that the approach has not yet been deployed in a live real-time crisis scenario. Future work should validate the system's timeliness and operational value by testing it in a live monitoring environment, for example during an ongoing crisis or in a drill with emergency responders. The study was also limited by the input data related to tweets collected during the earthquake. Future research could incorporate multimodal data sources to improve the robustness of anomaly detection. Additionally, this study was limited to a single crisis event (i.e., tweets from the February 2023 Turkey–Syria earthquake). Techniques such as data augmentation or transfer learning could be employed to handle varying conditions, and to reduce any dataset-specific bias. Another limitation is that this analysis was limited to English-language tweets. This focus may restrict the cross-cultural applicability of the findings. Future work could, therefore, extend the study's approach to multilingual settings by utilizing language-diverse data. It is worth noting that the proposed sentiment classifier, although fine-tuned on the collected tweets, relies on a general-domain pretrained model. To better handle crisis-specific vocabulary (e.g., informal or urgent expressions), future work could explore domain-adaptive pretraining and the integration of crisis-specific sentiment resources or lexicons. Such domain-specific enhancements could further improve classification accuracy for nuanced content that a generic model might miss. Other limitations relate to the visualization and interpretability challenges associated with large-scale anomaly detection outcomes. Although the 95th percentile threshold theoretically flagged thousands of tweets as anomalous, practical visualizations were limited to manageable subsets (random 80 anomalies per method). Future research could explore adaptive thresholding or hierarchical anomaly detection to balance comprehensiveness with practical interpretability.

This study analyzes publicly available social media data, but it is important to acknowledge the ethical implications.

1. Bias: the language model used (BERT) may have inherited biases that could affect the fairness of sentiment analysis. Certain dialects or demographics might be misinterpreted, potentially leading to skewed results.
2. Privacy: all data analyzed were public tweets. However, the authors respected user privacy by working with anonymized datasets, and by adhering to platform policies. In any operational setting, safeguards must ensure that individual personal data is not misused, and that analyses remain aggregated and focused on public information needs.
3. False alarms: the authors caution that this anomaly system could trigger false positives. For instance, a surge in negative sentiment might be due to misinformation or social

panic that does not correspond to a real situation on the ground. Any alert should be verified and complemented by human analysis.

6. Conclusions

In this study, a transformer-based BERT model for sentiment analysis and autoencoders for anomaly detection outperformed traditional machine learning approaches in analyzing crisis-related tweets. Neural networks, with their ability to capture the contextual relationship within the text, make them an effective tool for crisis monitoring. Unlike traditional machine learning methods that require manual feature engineering, neural networks are capable of autonomously learning complex patterns from data. The potential of neural network methods to provide early warning indicators of sentiment changes can enable decision-makers and response teams to reach out proactively. By leveraging contextual understanding of sentiment, stakeholders can gain insights and facilitate more effective resource allocation and policymaking during crisis situations.

The neural network anomaly detection models demonstrated their effectiveness in recognizing deviations from usual sentiment patterns, which can serve as an important signal for emerging threats or unexpected public reactions. Anomaly detection based on LSTM with an attention mechanism showed effectiveness in detecting sentiment shifts, while the autoencoder-based model effectively captured anomalous sentiment patterns based on reconstruction errors. The combined use of these models increased the robustness of anomaly detection, providing an approach to understanding sentiment anomalies in social media datasets. This study demonstrated the potential of using neural networks in crisis-related events. Future research could use data from other sources like Facebook, Instagram, and TikTok. Integrating data from multiple sources could possibly provide a richer context for understanding crisis events.

Author Contributions: Conceptualization, J.K. and I.D.; methodology, J.K. and I.D.; software, J.K.; validation, I.D.; formal analysis, J.K. and I.D.; investigation, J.K.; resources, J.K.; data curation, J.K.; writing—original draft preparation, J.K.; writing—review and editing, I.D.; visualization, J.K.; supervision, I.D.; funding acquisition, I.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tamer, Z.; Demir, G.; Darıcı, S.; Pamučar, D. Understanding twitter in crisis: A roadmap for public sector decision makers with multi-criteria decision making. *Environ. Dev. Sustain.* **2025**, 1–37. [CrossRef]
2. Noor, N.; Okhai, R.; Jamal, T.B.; Kapucu, N.; Ge, Y.G.; Hasan, S. Social-media-based crisis communication: Assessing the engagement of local agencies in Twitter during Hurricane Irma. *Int. J. Inf. Manag. Data Insights* **2024**, *4*, 100236. [CrossRef]
3. Karimiziarani, M.; Moradkhani, H. Social response and Disaster management: Insights from twitter data Assimilation on Hurricane Ian. *Int. J. Disaster Risk Reduct.* **2023**, *95*, 103865. [CrossRef]
4. Kanungo, S.; Jain, S. Hybrid Deep Neural Network G-LSTM for Sentiment Analysis on Twitter: A Novel Approach to Disaster Management. *Ingénierie Des Systèmes D'information* **2023**, *28*, 1565–1575. [CrossRef]
5. Kumar, S.; Khan, M.B.; Hasanat, M.H.; Saudagar, A.K.; Al Tameem, A.; Al Khathami, M. An anomaly detection framework for twitter data. *Appl. Sci.* **2022**, *12*, 11059. [CrossRef]
6. Liu, D.; Zhao, Y.; Xu, H.; Sun, Y.; Pei, D.; Luo, J.; Jing, X.; Feng, M. Opprentice: Towards practical and automatic anomaly detection through machine learning. In Proceedings of the 2015 Internet Measurement Conference, Tokyo, Japan, 28–30 October 2015; pp. 211–224. [CrossRef]

7. Elmrabit, N.; Zhou, F.; Li, F.; Zhou, H. Evaluation of Machine Learning Algorithms for Anomaly Detection. In Proceedings of the 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Dublin, Ireland, 15–19 June 2020; pp. 1–8. [CrossRef]
8. Rahman, M.S.; Halder, S.; Uddin, M.A.; Acharjee, U.K. An efficient hybrid system for anomaly detection in social networks. *Cybersecurity* **2021**, *4*, 10. [CrossRef]
9. Steuber, F.; Schneider, S.; Schneider, J.A.; Rodosek, G.D. Real-Time Anomaly Detection and Popularity Prediction for Emerging Events on Twitter. In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Kusadasi, Turkiye, 6–9 November 2023; pp. 300–304. [CrossRef]
10. Sufi, F.K.; Alsulami, M. Automated Multidimensional Analysis of Global Events With Entity Detection, Sentiment Analysis and Anomaly Detection. *IEEE Access* **2021**, *9*, 152449–152460. [CrossRef]
11. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
12. Koroteev, M.V. BERT: A review of applications in natural language processing and understanding. *arXiv* **2021**. [CrossRef]
13. Lee, M.C.; Lin, J.C.; Gran, E.G. SALAD: Self-Adaptive Lightweight Anomaly Detection for Real-time Recurrent Time Series. In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; pp. 344–349. [CrossRef]
14. Atkinson, O.; Bhardwaj, A.; Englert, C.; Ngairangbam, V.S.; Spannowsky, M. Anomaly detection with convolutional Graph Neural Networks. *J. High Energy Phys.* **2021**, *2021*, 80. [CrossRef]
15. Das, R.; Singh, T.D. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.* **2023**, *55*, 270. [CrossRef]
16. Katalinić, J.; Dunđer, I.; Seljan, S. Unraveling the Nuclear Debate: Insights Through Clustering of Tweets. *Electronics* **2024**, *13*, 4159. [CrossRef]
17. Katalinić, J.; Dunđer, I.; Seljan, S. Polarizing Topics on Twitter in the 2022 United States Elections. *Information* **2023**, *14*, 609. [CrossRef]
18. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
19. Shetty, N.P.; Bijalwan, Y.; Chaudhari, P.; Shetty, J.; Muniyal, B. Disaster assessment from social media using multimodal deep learning. *Multimed. Tools Appl.* **2024**, *83*, 17–23. [CrossRef]
20. Fernandez, G.; Suresh-Babu, S.; Vito, D. Mapping Infodemic Responses: A Geospatial Analysis of COVID-19 Discourse on Twitter in Italy. *Int. J. Environ. Res. Public Health* **2025**, *22*, 668. [CrossRef]
21. Suhasini, M.; Srinivasu, B. Emotion Detection Framework for Twitter Data Using Supervised Classifiers. *Adv. Intell. Syst. Comput.* **2020**, *1079*, 565–576. [CrossRef]
22. Jayakody, J.P.U.S.D.; Kumara, B.T.G.S. Sentiment analysis on product reviews on twitter using Machine Learning Approaches. In Proceedings of the 2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 7–8 December 2021; pp. 1056–1061. [CrossRef]
23. Tan, K.L.; Lee, C.P.; Anbananthen, K.S.M.; Lim, K.M. RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. *IEEE Access* **2022**, *10*, 21517–21525. [CrossRef]
24. Stojanovski, D.; Strezoski, G.; Madjarov, G.; Dimitrovski, I.; Chorbev, I. Deep neural network architecture for sentiment analysis and emotion identification of Twitter messages. *Multimed. Tools Appl.* **2018**, *77*, 32213–32242. [CrossRef]
25. Myint, P.Y.; Lo, S.L.; Zhang, Y. Unveiling the dynamics of crisis events: Sentiment and emotion analysis via multi-task learning with attention mechanism and subject-based intent prediction. *Inf. Process. Manag.* **2024**, *61*, 103695. [CrossRef]
26. Nguyen, D.Q.; Vu, T.; Nguyen, A.T. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 9–14. [CrossRef]
27. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lwei, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**. [CrossRef]
28. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1724–1734. [CrossRef]
29. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
30. Peng, H.; Zhang, R.; Li, S.; Cao, Y.; Pan, S.; Yu, P.S. Reinforced, Incremental and Cross-Lingual Event Detection From Social Messages. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 980–998. [CrossRef]

31. Patel, K.; Hoeber, O.; Hamilton, H. Real-Time Sentiment-Based Anomaly Detection in Twitter Data Streams. In Proceedings of the 28th Canadian Conference on Artificial Intelligence (Canadian AI 2015), Halifax, NS, Canada, 2–5 June 2015; pp. 196–203. [CrossRef]
32. Roy, A.; Shu, J.; Li, J.; Yang, C.; Elshocht, O.; Smeets, J.; Li, P. GAD-NR: Graph Anomaly Detection via Neighborhood Reconstruction. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), Merida, Mexico, 4–8 March 2024; pp. 576–585. [CrossRef]
33. Wong, L.; Liu, D.; Berti-Equille, L.; Alnegheimish, S.; Veeramachaneni, K. AER: Auto-encoder with regression for time series anomaly detection. In Proceedings of the 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 17–20 December 2022; pp. 1152–1161. [CrossRef]
34. Do, J.S.; Kareem, A.B.; Hur, J.W. LSTM-Autoencoder for Vibration Anomaly Detection in Vertical Carousel Storage and Retrieval System (VCSRS). *Sensors* **2023**, *23*, 1009. [CrossRef] [PubMed]
35. Kaggle. Available online: https://www.kaggle.com/datasets/swaptr/turkey-earthquake-tweets (accessed on 8 March 2025).
36. Sahoo, A.; Chanda, R.; Das, N.; Sadhukhan, B. Comparative Analysis of BERT Models for Sentiment Analysis on Twitter Data. In Proceedings of the 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, India, 17–19 August 2023; pp. 658–663. [CrossRef]
37. Hickman, L.; Thapa, S.; Tay, L.; Cao, M.; Srinivasan, P. Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations. *Organ. Res. Methods* **2022**, *25*, 114–146. [CrossRef]
38. Al-Khafaji, H.K.; Habeeb, A.T. Efficient Algorithms for Preprocessing and Stemming of Tweets in a Sentiment Analysis System. *IOSR J. Comput. Eng. (IOSR-JCE)* **2017**, *19*, 44–50. [CrossRef]
39. Roy, D.; Mitra, M.; Ganguly, D. To Clean or Not to Clean: Document Preprocessing and Reproducibility. *J. Data Inf. Qual. (JDIQ)* **2018**, *10*, 18. [CrossRef]
40. Lakhanpal, S.; Gupta, A.; Agrawal, R. Leveraging Explainable AI to Analyze Researchers' Aspect-Based Sentiment About ChatGPT. In Proceedings of the 15th International Conference on Intelligent Human Computer Interaction (IHCI 2023), Daegu, Republic of Korea, 8–10 November 2023; pp. 281–290. [CrossRef]
41. Hussain, Z.; Binz, M.; Mata, R.; Wulff, D.U. A tutorial on open-source large language models for behavioral science. *Behav. Res.* **2024**, *56*, 8214–8237. [CrossRef] [PubMed]
42. Siegel, J.W. Optimal Approximation Rates for Deep ReLU Neural Networks on Sobolev and Besov Spaces. *J. Mach. Learn. Res.* **2023**, *24*, 1–52. [CrossRef]
43. Foorthuis, R. On the nature and types of anomalies: A review of deviations in data. *Int. J. Data Sci. Anal.* **2021**, *12*, 297–331. [CrossRef]
44. Cai, Y.; Chen, H.; Cheng, K.-T. Rethinking autoencoders for medical anomaly detection from a theoretical perspective. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024), Marrakesh, Morocco, 6–10 October 2024; pp. 544–554. [CrossRef]
45. Wang, Y.; Du, X.; Lu, Z.; Duan, Q.; Wu, J. Improved LSTM-Based Time-Series Anomaly Detection in Rail Transit Operation Environments. *IEEE Trans. Ind. Inform.* **2022**, *18*, 9027–9036. [CrossRef]
46. Premasudha, B.G.; Rampalli, V. A Comparative Study of Logistic Regression, Support Vector Machines, and LSTM Networks for Sentiment Classification in Academic Reviews. In Proceedings of the 2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC), Davangere, India, 24–25 October 2024; pp. 1–11. [CrossRef]
47. Jahan, I.; Islam, M.N.; Hasan, M.M.; Siddiky, M.R. Comparative analysis of machine learning algorithms for sentiment classification in social media text. *World J. Adv. Res. Rev.* **2024**, *23*, 2842–2852. [CrossRef]
48. Jamil, M.L.; Pais, S.; Cordeiro, J.; Dias, G. Detection of extreme sentiments on social networks with BERT. *Soc. Netw. Anal. Min.* **2022**, *12*, 55. [CrossRef]
49. Santoro, D.; Ciano, T.; Ferrara, M. A comparison between machine and deep learning models on high stationarity data. *Sci. Rep.* **2024**, *14*, 19409–19413. [CrossRef] [PubMed]
50. Zwetsloot, I.M.; Jones-Farmer, L.A.; Woodall, W.H. Monitoring univariate processes using control charts: Some practical issues and advice. *Qual. Eng.* **2024**, *36*, 487–499. [CrossRef]
51. Rezk, M.; Elmadany, N.; Hamad, R.K.; Badran, E.F. Categorizing Crises From Social Media Feeds via Multimodal Channel Attention. *IEEE Access* **2023**, *11*, 72037–72049. [CrossRef]
52. Alhashmi, S.M.; Khedr, A.M.; Arif, I.; El-Bannany, M. Using a Hybrid-Classification Method to Analyze Twitter Data During Critical Events. *IEEE Access* **2021**, *9*, 141023–141035. [CrossRef]