

A Hybrid Machine Learning Framework for Sentiment-Based Early Crisis Detection in News Streams

Sena Nur Çulcu¹ and Canan Koç^{1*}

¹ Department of Software Engineering, Faculty of Engineering, Fırat University, Elazığ, Türkiye

*Corresponding author: canan.koc@firat.edu.tr

ABSTRACT

The rapid growth of digital news sources has made the automatic analysis of news content increasingly important for the early identification of potential societal crises. In this study, an integrated approach is proposed to classify the sentiment of news headlines and to support the identification of potential crisis periods by monitoring temporal changes in the intensity of negative sentiment. News headlines from the News Category Dataset V3 were categorized into three sentiment classes—positive, neutral, and negative—using the VADER sentiment analysis method. For text representation, TF-IDF features, which capture the statistical importance of words, and MiniLM embeddings, which model semantic and contextual relationships, were employed. A hybrid representation was constructed by combining these two feature groups, and several machine learning models were comparatively evaluated on this representation. The best-performing model, a hybrid Linear SVM combining TF-IDF and MiniLM features, achieved 91.22% accuracy and a Macro-F1 score of 0.9099 in predicting the VADER-generated sentiment labels. Furthermore, when comparing VADER labels with independent human evaluations, the model achieved 86.67% accuracy, a Macro-F1 score of 0.8659, and a Cohen's Kappa value of 0.7975. The resulting hybrid representation was then integrated into the proposed temporal crisis assessment framework, enabling sentiment classification outputs to monitor unusual changes in negative sentiment over time. The proposed approach not only classifies news headlines by sentiment but also enables analysis of temporal changes in the proportion of negative sentiment. By integrating the statistical strength of TF-IDF with the contextual representation capability of MiniLM, the proposed framework provides a high-performing, lightweight, and practical solution. It is anticipated that this study will contribute to decision support systems for monitoring news streams and facilitating the early assessment of potential crisis trends.

Keywords: Crisis detection; sentiment analysis; news headlines; TF-IDF; MiniLM; Linear SVM; machine learning.

1. INTRODUCTION

The rapid growth of content generated on news platforms and digital media has enabled crisis-related information to reach large audiences in a very short time. Consequently, the automatic analysis of news content has become important not only for news classification and sentiment analysis but also for the early identification of potential crisis indicators. However, most existing studies primarily focus on categorizing news articles or social media posts, or on detecting crisis events after they have already occurred. In contrast, studies investigating temporal changes in negative sentiment within news headlines as an early indicator

of potential crises remain limited. Although many state-of-the-art methods achieve high classification performance, they often require large-scale datasets, involve substantial computational costs, and rely on complex model architectures, making them less suitable for real-time applications. Furthermore, the limited interpretability of their decision-making processes represents a significant drawback in applications where reliability is critical, such as early warning systems. Effective crisis management requires more than merely identifying events after they occur. Detecting crisis signals at their early stages can enable decision-makers to respond promptly and mitigate potential adverse impacts. In this study, the term “crisis” is used in a broad, domain-independent sense to refer to emerging events or periods that may produce an unusual increase in negative sentiment within news streams. Accordingly, the proposed framework is not designed to classify a specific crisis type. Rather, it aims to identify general early-warning signals that may be associated with events such as natural disasters, economic or financial disruptions, political or social instability, and public health emergencies. The framework therefore indicates when an abnormal negative shift occurs in the news stream, while determining the specific nature of the underlying crisis requires subsequent contextual or domain-specific analysis. Therefore, continuously monitoring temporal changes in the intensity of negative sentiment expressed in news headlines can provide valuable information for the early detection of potential crises.

Motivated by these challenges, this study analyzes the proportion of negative sentiment over specific time intervals using sentiment scores derived from news headlines and evaluates their temporal variations through a dynamic thresholding approach. In this way, an interpretable and computationally efficient early warning mechanism is proposed to facilitate the early identification of potential crisis signals. Unlike existing studies, the proposed approach considers temporal fluctuations in negative sentiment expressed in news headlines as an early indicator of emerging crises, thereby offering a novel perspective on crisis detection.

Motivated by this perspective, this study analyzes weekly negative sentiment ratios derived from news headlines and evaluates substantial increases above the expected trend using a percentile-based thresholding approach. In this way, an interpretable and computationally efficient mechanism is proposed to support the identification of unusual increases in negative sentiment within the news stream as potential crisis candidates. From this perspective, the study aims to provide a distinct approach from existing research by treating temporal increases in negative sentiment in news headlines as potential indicators of emerging crisis periods.

1.1 Motivation

Previous studies have employed a variety of machine learning and deep learning approaches for news classification, sentiment analysis, and crisis-related content identification. Hamza et al. [1] combined GloVe word embeddings with a hybrid CNN–LSTM model, achieving over 91% accuracy on the AG News dataset and approximately 61% on the News Category Dataset V3. Zare et al. [2] integrated TF-IDF features with BiLSTM representations using an attention mechanism, obtaining accuracies of 94.10% on AG News and 89.10% on the 20 Newsgroups dataset. Aprilio et al. [3] combined TF-IDF and BERT features and achieved a test accuracy of 80.02%. The discrepancy between training and test performance suggests that the model exhibited limited generalization.

Zhang and Hu [4] combined BERT representations with a hypergraph-based variational framework and a capsule network, reporting accuracies ranging from 61.22% to 86.82%, depending on the dataset. Chhetri et al. [5] compared Transformer-based models for fake news and satire classification, achieving 87.58% accuracy with MiniLM and 95.42% ROC-AUC with RoBERTa. Although these studies demonstrate that contextual representations can substantially improve classification performance, they generally rely on computationally intensive and architecturally complex models.

More recently, large language models (LLMs) have emerged as an alternative paradigm for sentiment analysis, particularly through zero-shot and few-shot prompting strategies. Zhang et al. [6] systematically evaluated LLMs across multiple sentiment analysis tasks and datasets, showing that although LLMs perform satisfactorily on relatively simple sentiment tasks, their performance may decline on more complex sentiment phenomena; nevertheless, they demonstrated notable advantages in few-shot settings where labeled data are limited. In the news domain, Kuila and Sarkar [7] investigated zero-shot and few-shot LLM-based sentiment classification of political entities in news articles and reported that in-context learning substantially improved sentiment prediction performance. These findings demonstrate the growing potential of LLM-based sentiment analysis without conventional task-specific training. However, such approaches rely on substantially larger generative models and inference-oriented prompting, whereas the present study focuses on a lightweight hybrid TF-IDF–MiniLM representation coupled with Linear SVM, with the additional objective of continuously transforming sentiment predictions into temporal indicators for early crisis assessment.

Focusing on disaster-related content, Dharrao et al. [8] employed BERT embeddings combined with a CNN architecture, achieving 83% accuracy and 80% F1-score. Yicheng et al. [9] developed a BART–LSTM-based model to classify disaster-related tweets into 12 categories, reporting Micro-F1 and Macro-F1 scores of 0.8871 and 0.8764, respectively. However, these studies primarily focused on classifying disaster-related content or disaster types and did not investigate temporal variations in negative sentiment as an indicator of potential crises.

Jahan et al. [10] compared several machine learning algorithms for classifying social media texts into positive, negative, and neutral sentiment categories, with Logistic Regression and SVM achieving the highest accuracy of 86.22%. Peng et al. [11] evaluated 13 early warning indicators for large-scale emergencies using a genetic algorithm-optimized backpropagation neural network, obtaining an approximate prediction accuracy of 85%. Dui and Hu [12] integrated textual, audio, and visual features through an attention mechanism and reported class-specific accuracies of 78.2%, 81.5%, and 80.5% for the Sad, Happy, and Neutral categories, respectively, on the IEMOCAP dataset. Nevertheless, the approach proposed by Peng et al. relies on numerous predefined indicators, whereas the method of Dui and Hu requires multiple data modalities.

Most of the existing studies focus on classifying texts into news, sentiment, or disaster-related categories without integrating the resulting sentiment information into a temporal crisis assessment framework. In contrast, the proposed study labels news headlines as positive, neutral, or negative using the VADER sentiment analysis method and evaluates various machine learning models using enhanced TF-IDF features and MiniLM embeddings. The hybrid Linear SVM model achieved the best performance, reaching an accuracy of 91.22% and an F1-score of 90.99%.

Kahraman and Önder [13] temporally and thematically analyzed sentiment changes in social media posts related to the collapse of Signature Bank using VADER and FinBERT. In contrast, the present study differs from [13] by using general news headlines, not being limited to a single predefined crisis, and automatically identifying unusual increases in weekly negative sentiment ratios as potential crisis candidates through a moving-average and percentile-based thresholding approach.

The proposed approach provides a more comprehensive text representation by combining traditional statistical text features with contextual MiniLM embeddings, offering a richer representation than methods relying solely on TF-IDF or deep learning-based representations. The experimental results indicate that the proposed method achieves higher classification performance than many existing studies on fake news detection, disaster-related text classification, and sentiment analysis. However, since the datasets and classification tasks differ across studies, these numerical comparisons should not be interpreted as direct evidence of model superiority. Instead, they demonstrate the strong and competitive performance of the proposed approach within its specific application domain. The primary contribution of this study lies in integrating high-accuracy sentiment classification with a framework that evaluates temporal changes in negative sentiment in news headlines as an indicator of potential crises. A summary of the related literature is presented in Table 1.

TABLE 1. Literature Review

Ref.	Study	Dataset / Domain	Feature Extraction	Method	Sentiment Labeling	Crisis / Temporal Analysis	Best Result
[1]	Hamza et al. (2026)	AG News; News Category Dataset V3	Pre-trained embeddings	Hybrid neural network	Category-based	No	AG News >91%; News Category Dataset V3 ≈61%
[2]	Zare (2025)	20 Newsgroups; AG News	TF-IDF + BiLSTM	Attention-guided fusion	Category-based	No	89.1% / 94.1%
[3]	Aprilio et al. (2025)	CISI	TF-IDF (SVD) + BERT	Fully connected neural network	Binary relevance	No	80.02%
[4]	Zhang and Hu (2024)	CHIP-CTC; N15News	TF-IDF + semantic fusion	Fusion module	No	No	86.82% / 72.21% / 61.22%
[5]	Chhetri et al. (2026)	Satire / fake news	MiniLM embeddings	Lightweight Transformer	No	No	87.58%
[8]	Dharrao et al. (2024)	Disaster-related tweets	BERT embeddings	CNN + RMSProp	No	No	83.0%
[9]	Yicheng et al. (2026)	Multilingual disaster-related tweets	BART	BART + LSTM	No	No	Micro-F1 = 0.8871
[10]	Jahan et al. (2024)	100,000 tweets	TF-IDF	LR, SVM, RF, NB, GB	Three-class sentiment	No	86.22%
[11]	Peng et al. (2021)	Large-scale emergencies	13 indicators	GA-BP neural network	No	Yes (early warning)	≈85%
[12]	Dui and Hu (2024)	IEMOCAP	Multimodal features + attention	Multimodal sentiment classification	Three-class	Early-warning proposal	81.5% (maximum class-specific accuracy)

1.2 Contribution

The primary contribution of this study is the development of an integrated approach for determining the sentiment of news headlines and supporting the identification of potential crisis periods by analyzing temporal variations in the intensity of negative sentiment. To this end, news headlines were labeled as positive, neutral, or negative using the VADER sentiment analysis method. For text representation, TF-IDF-based features, which capture the statistical importance of words, were combined with MiniLM embeddings, which encode semantic and contextual relationships among words. This hybrid representation provides a more comprehensive feature space than approaches relying exclusively on word-frequency statistics or contextual embeddings.

A comparative evaluation of multiple machine learning models was conducted to assess the effectiveness of the proposed hybrid representation. Beyond sentiment classification, the framework integrates the resulting sentiment information with temporal analysis of negative sentiment ratios to identify unusual increases as potential crisis candidates. In addition to its high classification performance, the proposed approach provides a lightweight and practically applicable model architecture. The study goes beyond merely classifying news headlines into sentiment categories by enabling temporal monitoring of sudden changes in negative sentiment ratios. In this context, building on the previously established relationship between increases in negative sentiment and crisis periods, the study aims to evaluate this relationship within a threshold-based mechanism for identifying potential crisis candidates from general news headlines.

2. MATERIALS AND METHODS

2.1 Dataset

In this study, 209,520 English-language news headlines from the News Category Dataset V3, covering the period from January 28, 2012 to September 23, 2022, were analyzed. The headlines were labeled as positive, negative, or neutral using the VADER method. The dataset consists of English-language news content and is provided in JSON Lines format, containing fields such as news headline, news category, short description, author information, URL, and publication date. News headlines constitute the primary and sole textual input of the proposed approach. To generate the target variable for sentiment classification, the headlines were labeled using VADER into one of three sentiment classes: positive, neutral, or negative. These labels were subsequently used to train and evaluate machine learning models and classification models based on hybrid feature extraction. In addition, the publication dates of the news articles were utilized to analyze temporal variations in the proportion of negative sentiment and to identify potential crisis signals.

The main characteristics of the dataset used in this study are presented in Table 2.

TABLE 2. Features of Dataset

Feature	Description
Dataset	News Category Dataset V3
Total Number of Headlines	209,520
Date Range	January 28, 2012 – September 23, 2022
Negative	58,890
Neutral	81,368
Positive	69,262
Training Set	167,616
Test Set	41,904
Training/Test Split	80% / 20%
Text Field	Headline
Sentiment Classes	Positive, Neutral, Negative

After preprocessing, a total of 209,520 news headlines were included in the analysis. Based on the VADER-generated sentiment labels, 58,890 headlines were classified as negative, 81,368 as neutral, and 69,262 as positive. The publication dates of the analyzed news items ranged from January 28, 2012 to September 23, 2022. The dataset was divided into 167,616 training samples and 41,904 test samples using an 80%/20% stratified split.

2.2 Data Preprocessing

Data preprocessing is a critical step in text mining and sentiment analysis, as it directly influences the performance of classification models. Therefore, before the news headlines were included in the classification pipeline, a series of preprocessing operations was performed. First, all URLs embedded in the headlines were removed. Subsequently, all text was converted to lowercase to eliminate inconsistencies arising from case sensitivity. Punctuation marks, special characters, and other symbols that did not contribute meaningful information to the analysis were removed, and multiple consecutive whitespace characters were normalized into a single space. These preprocessing steps transformed the news headlines into a standardized textual format.

After preprocessing, records with empty or missing text were excluded from the dataset. The cleaned news headlines were stored in the `headline_clean` field and subsequently used for feature extraction. In addition, whenever the `short_description` field was available, it was concatenated with the corresponding news headline to create a richer textual representation. Since the VADER method considers linguistic cues such as capitalization and punctuation, sentiment labels were generated directly from the raw headline field without applying any text preprocessing. In contrast, the preprocessed `headline_clean` field was used for TF-IDF- and MiniLM-based feature extraction.

To train and evaluate the classification models, the dataset was divided into training and test sets. Specifically, 80% of the data were allocated to training, and the remaining 20% to testing. To preserve the class distribution across both subsets, stratified sampling was employed. This ensured that similar class proportions were maintained in the training and test sets, leading to a more reliable and unbiased evaluation of model performance.



2.3 Sentiment Labeling with VADER

In this study, the VADER method was applied to the raw, unprocessed headline field to generate sentiment labels for the news headlines.

Following the data preprocessing stage, the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis method was employed to determine the sentiment class of each news headline. VADER is a lexicon- and rule-based sentiment analysis approach that has been widely adopted in the literature due to its effectiveness in analyzing short texts, social media content, and user-generated comments. The method evaluates the sentiment orientation of words and expressions within a text and produces positive, negative, and neutral sentiment scores, along with an overall sentiment polarity score (compound score) [14].

In the VADER framework, each word is assigned a predefined sentiment weight, and these weights are combined using a set of linguistic and grammatical rules to estimate the overall sentiment of the text. The method also considers textual characteristics such as exclamation marks, capitalization, degree modifiers, and abbreviations, enabling a more realistic estimation of sentiment scores. Furthermore, numerous studies have reported that VADER achieves strong performance on social media data, short texts, and online content [14,15]. Owing to its effectiveness in distinguishing between positive, negative, and neutral sentiment classes, VADER is widely regarded as a reliable approach for sentiment analysis tasks [15].

In this study, the compound sentiment score generated by VADER was used to determine the sentiment of each news headline. The compound score ranges from -1 to +1, representing the overall sentiment polarity of the text. The resulting compound scores were categorized according to the standard VADER threshold values. Headlines with a compound score of 0.05 or higher were labeled as positive, those with a score of -0.05 or lower were labeled as negative, and those with scores between -0.05 and 0.05 were labeled as neutral. Consequently, all news headlines in the dataset were assigned to one of three sentiment classes: positive, neutral, or negative.

The generated sentiment labels were subsequently used as the target variable for the machine learning and hybrid classification models. By combining these sentiment labels with statistical and contextual textual features, a comprehensive classification dataset was constructed, forming the foundation of the proposed crisis signal detection framework. The sentiment classification thresholds used in this study are presented in Table 3.

TABLE 3. VADER Sentiment Classification Thresholds

Compound Score	Sentiment Label
compound ≥ 0.05	Positive
-0.05 < compound < 0.05	Neutral
compound ≤ -0.05	Negative

2.3.1 Construction of the Human-Annotated Validation Set

To assess the reliability of the sentiment labels automatically generated by VADER, 300 news headlines were randomly selected from the dataset while preserving the VADER class distribution. The selected headlines were independently annotated as positive, neutral, or negative by two annotators who were blinded to both the VADER-generated labels and compound sentiment scores. In the initial independent annotation, the two annotators assigned the same class to 267 of the 300 headlines, corresponding to an observed agreement of 89.00%. To assess inter-annotator agreement beyond that expected by chance, Cohen's Kappa coefficient was calculated and found to be 0.832. The 33 headlines for which the annotators assigned different labels were subsequently re-evaluated, and the final human annotations were established through consensus. Agreement between the VADER-generated labels and the final human annotations was evaluated using accuracy, Macro-F1, and Cohen's Kappa.

2.4 Feature Extraction

2.4.1 TF-IDF-Based Feature Extraction

Since machine learning algorithms cannot process raw textual data directly, textual information must first be transformed into numerical feature vectors. To achieve this, the Term Frequency–Inverse Document Frequency (TF-IDF) method, one of the most widely used feature extraction techniques in natural language processing and text classification, was employed. TF-IDF determines the importance of a term within a specific document while simultaneously considering its distribution across the entire document collection. As a result, the influence of frequently occurring but less discriminative words is reduced, whereas the weights of terms that are more representative of individual documents are increased. Previous studies have demonstrated that TF-IDF is highly effective for text classification tasks and has been extensively applied in the classification of news articles [16,17].

TF-IDF is one of the fundamental techniques for representing textual documents as numerical feature vectors. In this approach, each term is treated independently, and documents are represented as high-dimensional vectors composed of term weights. The TF-IDF value of a term is computed as the product of its Term Frequency (TF), which measures how frequently the term appears within a document, and its Inverse Document Frequency (IDF), which reflects how widely the term is distributed across the entire document collection [16].

The mathematical formulation of the TF-IDF weighting scheme is presented in Equation (1).

$$TF - IDF(t, d) = TF(t, d) \times \left(\log \left(\frac{1+N}{1+DF(t)} \right) + 1 \right) \quad (1)$$

In Equation (1), TF-IDF(t,d) denotes the TF-IDF weight of term t in document d. TF(t,d) represents the term frequency, indicating the number of times term t appears in document d, whereas DF(t) denotes the number of documents containing that term. N represents the total number of documents in the dataset. The logarithmic weighting scheme employed in the equation reduces the influence of terms that occur frequently across many documents while increasing the importance of more discriminative terms. Consequently, TF-IDF enables a more informative representation of news headlines.

During the TF-IDF vectorization process, unigrams, bigrams, and trigrams were simultaneously employed to capture both individual words and multi-word expressions. Accordingly, the n-gram range was set to 1–3. No additional supervised feature-selection method, such as chi-square-based feature selection, was applied. The $\text{min_df}=2$ criterion served only as a frequency-based filtering step by excluding terms occurring in fewer than two training documents. After fitting the TF-IDF vectorizer exclusively on the training set, the resulting vocabulary contained 138,986 features. Consequently, each headline was represented as a 138,986-dimensional sparse TF-IDF vector. The resulting TF-IDF matrices had dimensions of $167,616 \times 138,986$ for the training set and $41,904 \times 138,986$ for the test set. The test set was transformed using the vocabulary learned exclusively from the training data to prevent information leakage. To reduce noise from infrequent terms, the minimum document frequency (min_df) was set to 2, and terms appearing in more than 90% of documents were excluded due to their limited discriminative power. Furthermore, to maintain a sufficiently expressive feature space, the maximum number of features (max_features) was limited to 250,000, and sublinear term frequency scaling was applied. To prevent data leakage, the TF-IDF vectorizer was fitted (fit) exclusively on the news headlines in the training set, while the headlines in the test set were only transformed (transform) using the vocabulary and IDF weights learned from the training data. Thus, no information from the test set was used during the construction of the TF-IDF feature space.

The resulting TF-IDF feature vectors were directly used to train the Logistic Regression, Multinomial Naïve Bayes, Linear Support Vector Machine (Linear SVM), and Voting Ensemble classifiers. In addition, the TF-IDF-derived statistical features were combined with the semantic representations generated by the Transformer-based MiniLM model, forming one of the key components of the proposed hybrid classification framework.

2.4.2 MiniLM Transformer-Based Feature Extraction

In recent years, Transformer-based language models have become widely adopted in natural language processing for obtaining semantic representations of textual data. While traditional feature extraction methods primarily rely on the statistical distribution of words within documents, Transformer models learn contextual relationships among words, enabling the generation of richer semantic representations. Consequently, Transformer-based embedding techniques have been extensively applied in sentiment analysis, text classification, and information retrieval tasks [18,19].

In this study, semantic features were extracted using the all-MiniLM-L6-v2 model available in the Sentence Transformers library. MiniLM is a lightweight language model developed through knowledge distillation from large Transformer models, providing strong semantic representation capabilities with substantially lower computational cost [18]. Built on the Sentence-BERT (SBERT) architecture, the model transforms sentences and short texts into fixed-dimensional dense vectors, enabling more effective representation of semantic similarities between texts [19].

Although the TF-IDF approach effectively captures the statistical importance of words based on their frequency, it cannot adequately model semantic relationships between texts that convey similar meanings using different vocabulary. In contrast, the MiniLM model learns contextual relationships within the text, enabling semantically similar expressions to be represented by

neighboring vectors in the embedding space. This capability is particularly advantageous for short texts such as news headlines, where contextual information plays a crucial role in accurately identifying sentiment.

In the proposed framework, each news headline was encoded into a dense semantic vector using the MiniLM model. The resulting embeddings were subsequently combined with the TF-IDF feature vectors. In this way, both frequency-based statistical information and Transformer-based semantic representations were integrated into a unified feature space, forming the second key component of the proposed hybrid classification framework.

2.5 Classification Models

2.5.1 Logistic Regression

Logistic Regression (LR) is a widely used linear machine learning algorithm for classification tasks. The algorithm models the probabilistic relationship between the input features and the target variable, estimating the probability that a given instance belongs to a particular class. Owing to its effectiveness in handling high-dimensional datasets and its interpretable decision-making process, Logistic Regression has been extensively applied in text classification studies [20].

In text mining applications, Logistic Regression often achieves strong performance when combined with vector space representations such as TF-IDF. Hassan et al. [20] compared several machine learning algorithms and reported that Logistic Regression provides competitive performance for text classification tasks while effectively handling high-dimensional textual feature spaces. For this reason, Logistic Regression was included in this study as one of the baseline classification models using the TF-IDF feature representations extracted from the news headlines.

During training, balanced class weights were used to mitigate class imbalance, and the maximum number of iterations was set to 4,000. These parameter settings were intended to ensure stable convergence and enable the model to reach an optimal solution without encountering convergence issues.

2.5.2 Multinomial Naïve Bayes

Multinomial Naïve Bayes (MNB) is a probabilistic machine learning algorithm based on Bayes' theorem and is widely used in text classification and sentiment analysis applications. The algorithm assumes conditional independence among features and predicts the class of a given instance by estimating the posterior probability for each class. Owing to its simple architecture, low computational complexity, and strong performance on high-dimensional textual data, Multinomial Naïve Bayes has been extensively employed in natural language processing tasks [21].

The Naïve Bayes approach performs classification by utilizing the frequencies of words or terms within a document and is particularly well suited to TF-IDF-based feature representations. Yang [21] demonstrated that the Naïve Bayes algorithm provides interpretable results without requiring extensive hyperparameter tuning and achieves competitive performance in sentiment classification tasks. Furthermore, its low computational requirements make it highly efficient for large-scale text classification problems [21].

In this study, the Multinomial Naïve Bayes model was trained on TF-IDF feature vectors extracted from news headlines and included as one of the baseline models to evaluate the performance of the proposed hybrid classification framework.

2.5.3 Linear Support Vector Machine (Linear SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that aims to determine the optimal decision boundary for separating different classes. Owing to its strong performance on high-dimensional datasets, SVM has been widely applied in text classification and sentiment analysis tasks. The algorithm seeks to maximize the margin between classes, thereby improving its generalization capability and classification performance [22].

Because textual datasets are typically characterized by extremely high-dimensional feature spaces, the Linear Support Vector Machine (Linear SVM), which employs a linear kernel, is frequently preferred. Linear SVM is particularly effective when trained on high-dimensional TF-IDF feature representations while maintaining relatively low computational complexity during training. Moraes et al. [22] demonstrated that the SVM algorithm achieves high classification performance on textual data and produces more stable results than several alternative approaches in sentiment analysis.

In this study, the Linear SVM classifier was first trained on TF-IDF feature vectors extracted from news headlines and was included among the baseline models for performance comparison. Furthermore, within the proposed hybrid framework, Linear SVM was employed to classify the combined feature vectors obtained by integrating TF-IDF features with MiniLM embeddings. This choice was motivated by the proven effectiveness of Linear SVM in handling high-dimensional feature spaces and its strong performance across a wide range of text classification applications.

2.5.4 Voting Ensemble

Ensemble learning methods aim to improve predictive performance by combining the outputs of multiple classifiers to produce more robust and reliable predictions. By exploiting the complementary strengths of different learning algorithms, ensemble approaches seek to reduce the errors and biases that may arise from relying on a single model [23].

Among ensemble techniques, the Voting Ensemble method generates the final class prediction by aggregating the outputs of multiple classifiers. Previous studies have shown that ensemble methods can enhance classification performance by integrating the knowledge learned by different algorithms, thereby providing more stable and balanced predictions [23].

In this study, the Voting Ensemble model was constructed by combining the Logistic Regression, Multinomial Naïve Bayes, and Linear Support Vector Machine (Linear SVM) classifiers. The predictions from each base classifier were integrated via majority voting, and the resulting performance was compared with that of the individual models. This approach was adopted to leverage the complementary strengths of different classifiers and to establish a more robust and reliable decision-making framework.

2.6 Proposed Hybrid Approach

In this study, a hybrid classification framework is proposed for sentiment classification of news headlines by integrating TF-IDF-based statistical features with MiniLM-based semantic representations. The TF-IDF method provides an effective statistical

representation by considering both the frequency of terms within individual documents and their discriminative power across the entire corpus. However, it cannot explicitly capture semantic relationships between words. In contrast, the Transformer-based MiniLM model learns contextual information from text, generating semantically rich representations that effectively capture the underlying meaning of news headlines [18,19]. To leverage the complementary strengths of these two approaches, a hybrid feature-extraction framework was developed.

In the first stage of the proposed method, TF-IDF and MiniLM features were extracted independently from the preprocessed news headlines. The TF-IDF method was employed to derive statistical representations of the text, while the MiniLM model generated dense semantic embedding vectors capturing contextual relationships among words. Prior to feature concatenation, L2 normalization was applied separately to the TF-IDF and MiniLM features. The TF-IDF vectors were scaled to unit length using `TfidfVectorizer`, while the MiniLM embeddings were normalized using the model's built-in normalization option. This procedure reduced potential scale dominance arising from differences in the dimensionality and magnitude of the two feature representations. The normalized feature sets were then horizontally concatenated, with no additional normalization applied after concatenation. The TF-IDF representation consisted of 138,986 features, whereas the all-MiniLM-L6-v2 model generated a 384-dimensional semantic embedding for each headline. Accordingly, horizontal concatenation of the two representations resulted in a final hybrid feature vector of 139,370 dimensions (138,986 TF-IDF features + 384 MiniLM features) for each headline. The resulting hybrid vectors therefore simultaneously encode the statistical characteristics captured by TF-IDF and the semantic information represented by MiniLM. The overall architecture of the proposed hybrid sentiment classification framework is illustrated in Figure 1.

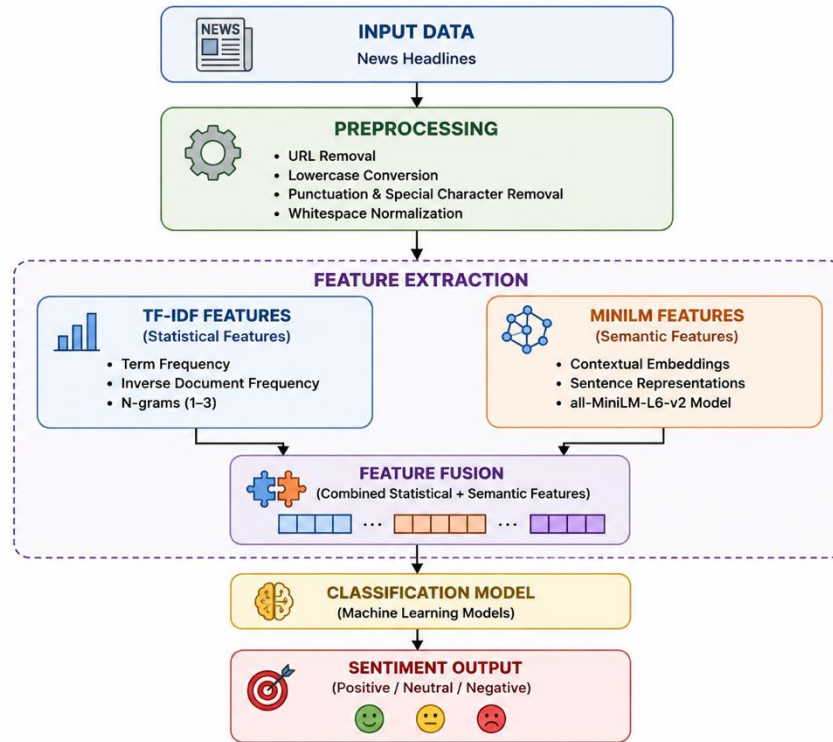


FIGURE 1. The proposed hybrid sentiment classification architecture based on the combination of TF-IDF and MiniLM features

As illustrated in Figure 1, the proposed hybrid architecture combines the statistical features extracted by TF-IDF with the semantic features generated by MiniLM to construct hybrid feature vectors. These hybrid feature vectors were then used to train a Linear Support Vector Machine (Linear SVM) classifier within the proposed framework. The performance of the proposed approach was subsequently compared with that of TF-IDF-based Logistic Regression, Multinomial Naïve Bayes, Linear SVM, and Voting Ensemble models. This comparison enabled a comprehensive evaluation of the contribution of integrating statistical and semantic features to sentiment classification performance.

2.7 Crisis Detection Mechanism

2.7.1 Negative Sentiment Ratio Analysis

In this study, sentiment labels derived from news headlines were analyzed over time to identify significant changes in negative sentiment intensity. Previous studies have indicated that sudden increases in negative sentiment observed in social media and news texts may be associated with crisis periods [13]. Accordingly, the temporal variation in negative sentiment ratios was examined to identify periods exhibiting unusual increases as potential crisis candidates.

Following sentiment analysis, each news headline was classified as positive, neutral, or negative. The news items were then grouped into weekly intervals based on their publication dates. For each week, both the total number of news headlines and the number of headlines classified as negative were calculated. The weekly negative sentiment ratio was computed by dividing the

number of negative headlines by the total number of headlines published during the corresponding week. This normalization reduced the influence of fluctuations in news volume across periods, yielding a more comparable indicator of negative sentiment over time.

The negative sentiment ratio was calculated using Equation (2).

$$R_t = \frac{N_t}{T_t} \quad (2)$$

In Equation (2), (R_t) denotes the negative sentiment ratio for week (t), (N_t) represents the number of news headlines classified as negative during that week, and (T_t) indicates the total number of news headlines published in the same period. Using the negative sentiment ratio, rather than the absolute number of negative news items, minimizes the influence of variations in the weekly volume of published news. Consequently, temporal changes in sentiment can be compared more consistently across time periods, providing a more reliable indicator for identifying potential crisis signals. Thus, an indicator is obtained for comparing temporal changes in sentiment and identifying unusual increases in negative sentiment intensity.

2.7.2 Percentile-Based Crisis Candidate Detection Using Positive Deviations

A percentile-based crisis candidate detection method was applied to identify significant upward changes in the negative sentiment ratio. First, the news articles were grouped on a weekly basis, and only weeks containing at least 20 news items were included in the analysis. For each week, the negative sentiment ratio was calculated as the number of negative news items divided by the total number of news items. Subsequently, a four-week moving average was used to estimate the expected trend of the negative sentiment ratio. As a crisis indicator, only positive deviations in which the observed negative sentiment ratio exceeded the expected trend were considered. This approach prevented decreases or downward deviations in the negative sentiment ratio from being incorrectly flagged as potential crisis candidates.

The deviation is defined in Equation (3) as follows:

$$d_t = R_t - \hat{R}_t \quad (3)$$

Here, d_t denotes the signed deviation in week t , R_t represents the observed weekly negative sentiment ratio, and $\hat{R}_t$ represents the expected negative sentiment ratio obtained using the four-week moving average. A positive value of d_t indicates that the observed negative sentiment ratio exceeds the expected trend. For crisis candidate detection, only positive deviations satisfying the condition $d_t > 0$ were considered.

$$Ct = \begin{cases} 1, & d_t > Q_{0.95}(D^+) \\ 0, & d_t \leq Q_{0.95}(D^+) \end{cases} \quad (4)$$

$$D^+ = \{d_t | d_t > 0\}$$

Here, C_t denotes whether week t is flagged as a crisis candidate, D^+ represents the set containing only positive deviations, and $Q_{0.95}(D^+)$ denotes the 95th percentile of the positive deviations. $C_t = 1$ indicates that the corresponding week is a crisis candidate, whereas $C_t = 0$ indicates that the specified threshold is not exceeded. This decision rule prevents downward changes from being interpreted as crisis signals.

A sensitivity analysis was conducted to evaluate the effect of the moving-average window length and percentile threshold selection on the obtained results. Accordingly, the moving-average window length was varied across three, four, and five weeks, while the threshold was set to the 90th, 95th, and 97.5th percentiles of the positive deviations. The four-week window was used as the baseline configuration because it represents approximately one month of historical trend and helps reduce the influence of short-term weekly fluctuations. The P95 threshold was selected because it provides a conservative criterion that considers only the upper 5% of positive deviations. The results obtained from different parameter combinations were compared in the Results section to assess the sensitivity of the proposed method to these parameter choices.

2.8 Performance Evaluation Criteria

The performance of the sentiment classification models developed in this study was evaluated using four widely adopted performance metrics: Accuracy, Precision, Recall, and F1-score. These evaluation metrics provide complementary perspectives on classification performance and are commonly employed in machine learning studies to assess the effectiveness of classification models [24]. Accuracy measures the overall classification performance of the model, whereas Precision evaluates the correctness of the classifier's positive predictions. Recall measures the model's ability to correctly identify actual positive instances, while the F1-score provides a balanced evaluation by considering both Precision and Recall simultaneously [24].

- **Accuracy**

Accuracy is defined as the proportion of correctly classified instances to the total number of instances in the dataset. It is calculated as shown in Equation (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

- **Precision**

Precision measures the proportion of instances the model predicts as positive that are actually positive. This metric measures the extent to which the model minimizes false-positive predictions. Precision is particularly important in classification tasks, where false-positive predictions have significant consequences, and is therefore widely regarded as a key performance metric for assessing classification models [24].

The Precision metric is calculated as shown in Equation (6).

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

- **Recall**

Recall indicates the proportion of actual positive samples in the dataset that are correctly identified by the model. It reflects the model's ability to capture samples belonging to the relevant class and is particularly important in applications where false-negative predictions are critical [24].

This metric is calculated as shown in Equation (7).

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

- **F1-Score**

The F1-score is a performance metric calculated as the harmonic mean of Precision and Recall. By simultaneously considering both false positive and false negative predictions, it provides a balanced evaluation of a classification model's performance. The F1-score is particularly valuable for datasets with imbalanced class distributions, where Accuracy alone may not provide a sufficiently reliable assessment of model performance [24].

The F1-score is calculated as shown in Equation (8).

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

In this study, the performance of Logistic Regression, Multinomial Naïve Bayes, Linear Support Vector Machine (Linear SVM), Voting Ensemble, and the proposed hybrid classification models was evaluated and compared using Accuracy, Precision, Recall, and F1-score. The experimental results obtained from these evaluations are presented and discussed in detail in the Results section.

In addition, to evaluate the contribution of the individual components of the proposed hybrid approach, three different feature representations—TF-IDF only, MiniLM only, and TF-IDF + MiniLM—were compared using the same Linear SVM classifier. To assess model generalizability, stratified 5-fold cross-validation was performed while preserving the class distribution, and the results were reported as mean $\pm$ standard deviation together with 95% confidence intervals. The statistical significance of the difference between the predictions of the TF-IDF and hybrid models on the same test samples was evaluated using the two-sided exact McNemar test. The statistical significance level was set at 0.05. The hyperparameters used in the experiments were based on standard configurations for the corresponding models and were kept fixed across the comparative experiments. An exhaustive hyperparameter optimization procedure was not performed, as the primary objective was to evaluate the contribution of the proposed feature representation under consistent experimental settings. This should be considered when interpreting the comparative results.

3. EXPERIMENTAL RESULTS

3.1 Performance Results of Machine Learning Models

The performance of the Logistic Regression, Multinomial Naïve Bayes, Linear Support Vector Machine (Linear SVM), Voting Ensemble, and the proposed hybrid classification models was evaluated using Accuracy, Precision, Recall, and F1-score. All models were trained and tested on the same training and test datasets to ensure a fair and consistent comparison. The results obtained were analyzed comparatively to assess the effectiveness of the proposed approach. The performance comparison of the machine learning-based classification models is presented in Table 4.

TABLE 4. Performance comparison of the machine learning-based classification models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9081	0.9107	0.9063	0.9082
Multinomial Naive Bayes	0.8033	0.8160	0.8144	0.8151
Linear SVM	0.9095	0.9085	0.9060	0.9071
Voting Ensemble	0.9065	0.9087	0.9069	0.9074
Proposed Method	0.9122	0.9107	0.9092	0.9099

As shown in Table 4, the proposed TF-IDF + MiniLM model achieved the highest classification accuracy among the evaluated models, with an accuracy of 0.9122. Among the baseline models using TF-IDF features, Linear SVM achieved an accuracy of 0.9095, followed by Logistic Regression with 0.9081 and the Voting Ensemble with 0.9065. Multinomial Naive Bayes achieved the lowest accuracy, with a value of 0.8033. These results indicate that combining TF-IDF features with MiniLM-based semantic representations provides a modest improvement in classification accuracy compared with the individual baseline models.

Among the individual models using only TF-IDF features, Linear SVM achieved the best performance, with an accuracy of 0.9095 and an F1-score of 0.9071. Logistic Regression showed comparable performance, achieving an accuracy of 0.9081. In contrast, Multinomial Naive Bayes yielded the lowest classification performance among the evaluated baseline models, with an accuracy of 0.8033. This result may be attributed to the conditional independence assumption underlying the Naive Bayes algorithm, which may limit its ability to adequately capture relationships among words in news headlines represented through TF-IDF features.

The Voting Ensemble model, which combined the predictions of Logistic Regression, Multinomial Naive Bayes, and Linear SVM through majority voting, achieved an accuracy of 0.9065. Ensemble methods do not necessarily outperform individual models in all cases. In this study, the comparatively lower performance of Multinomial Naive Bayes may have contributed to the Voting Ensemble model performing below Logistic Regression and Linear SVM. Overall, the results indicate that combining TF-IDF features with contextual features derived from MiniLM improves sentiment classification performance and contributes to more reliable results in the subsequent crisis detection process. Table 5 presents the results of the ablation analysis of different feature representations using Linear SVM.

TABLE 5. Ablation Analysis Results of Feature Representations Using Linear SVM

Feature Representation	Accuracy	Balanced Accuracy	Precision	Recall	Macro F1	Weighted F1	MCC	Kappa
TF-IDF	0.9095	0.9077	0.9085	0.9060	0.9071	0.9097	0.8630	0.8620
MiniLM	0.7091	0.6801	0.6824	0.6801	0.6807	0.7041	0.5420	0.5410
TF-IDF + MiniLM	0.9122	0.9092	0.9107	0.9092	0.9099	0.9121	0.8670	0.8660

Table 5 compares the effects of different feature representations using the same Linear SVM classifier. The model using only TF-IDF features achieved an accuracy of 90.95% and a Macro-F1 score of 90.71%, whereas the model using only MiniLM features achieved 70.91% and 68.07%, respectively. The proposed hybrid approach, which combines TF-IDF and MiniLM features, achieved the highest performance, with an accuracy of 91.22% and a Macro-F1 score of 90.99%. These findings indicate that MiniLM features alone do not provide sufficient classification performance; however, when combined with TF-IDF features, they provide a limited but positive contribution to classification performance. Table 6 presents the stratified 5-fold cross-validation results of the models.

TABLE 6. Stratified 5-Fold Cross-Validation Results of the Models

Feature Representation	Accuracy (Mean $\pm$ SD)	Accuracy 95% CI	Macro F1 (Mean $\pm$ SD)	Macro F1 95% CI
TF-IDF	0.9083 $\pm$ 0.0014	[0.9066–0.9100]	0.9060 $\pm$ 0.0015	[0.9041–0.9079]
MiniLM	0.6794 $\pm$ 0.0023	[0.6766–0.6823]	0.6800 $\pm$ 0.0023	[0.6772–0.6828]
TF-IDF + MiniLM	0.9104 $\pm$ 0.0008	[0.9094–0.9114]	0.9081 $\pm$ 0.0009	[0.9070–0.9092]

According to the stratified 5-fold cross-validation results presented in Table 6, the proposed TF-IDF + MiniLM model achieved the highest average performance, with an accuracy of 0.9104 ± 0.0008 and a Macro-F1 score of 0.9081 ± 0.0009 . The 95% confidence interval (CI) presented in the table represents the estimated range within which the mean performance value is expected to lie at the specified confidence level. The model using only TF-IDF features achieved an accuracy of 0.9083 ± 0.0014 and a Macro-F1 score of 0.9060 ± 0.0015 , whereas the model using only MiniLM features achieved 0.6794 ± 0.0023 and 0.6800 ± 0.0023 , respectively. The low standard deviations of the proposed hybrid model across the folds indicate that its performance remained stable across different data partitions. However, because the 95% confidence intervals of the hybrid and TF-IDF models partially overlap, the statistical significance of the difference between the two models was further evaluated using McNemar's test.

• Statistical Significance Analysis Between the Models

The statistical significance of the difference between the predictions of the TF-IDF and the proposed TF-IDF + MiniLM models on the same test set was evaluated using the two-sided exact McNemar test. The analysis identified 707 samples that were correctly classified by the TF-IDF model but incorrectly classified by the hybrid model, and 819 samples that

were incorrectly classified by the TF-IDF model but correctly classified by the hybrid model. Based on a total of 1,526 discordant predictions, the test yielded $p=0.0045$. Since this value is below the significance level of 0.05, the difference between the error patterns of the two models is statistically significant. The greater number of corrected predictions in favor of the hybrid model further supports that incorporating MiniLM features into the TF-IDF representation provides a modest but statistically significant improvement.

• **Confusion Matrix Analysis of the Proposed Model**

To evaluate the class-specific prediction performance of the proposed TF-IDF + MiniLM model in greater detail, the confusion matrix obtained on the test set was examined. The matrix illustrates the distribution of correct and incorrect predictions across the negative, neutral, and positive sentiment classes. The resulting confusion matrix is presented in Figure 2.

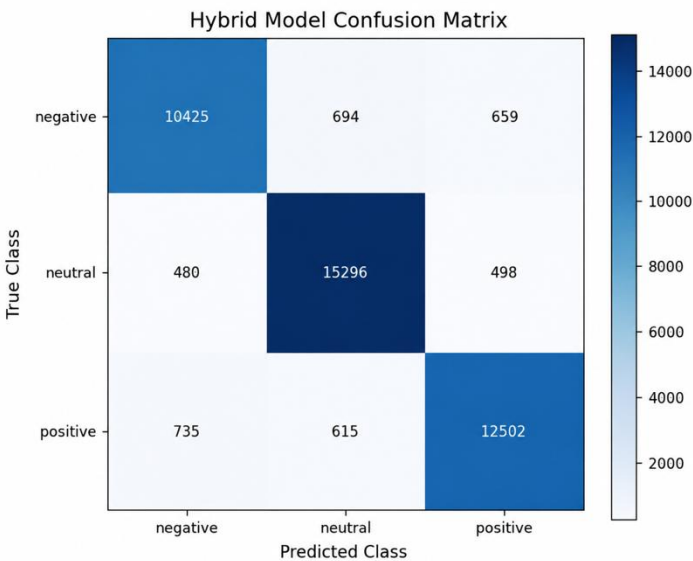


FIGURE 2. Confusion Matrix of the Proposed TF-IDF + MiniLM Model

Figure 2 presents the confusion matrix of the proposed TF-IDF + MiniLM model on the test set. The model correctly classified 10,425 of the 11,778 news headlines in the negative class, 15,296 of the 16,274 headlines in the neutral class, and 12,502 of the 13,852 headlines in the positive class. The highest number of correct classifications was obtained for the neutral class. Regarding misclassifications, 694 negative headlines were classified as neutral and 659 as positive; 480 neutral headlines were classified as negative and 498 as positive; and 735 positive headlines were classified as negative and 615 as neutral. These results indicate that the model achieved balanced classification performance across all three sentiment classes and was particularly effective in distinguishing the neutral class.



- **Validation of VADER Labels Against Human Annotations**

To examine the agreement between the sentiment labels automatically generated by VADER and human assessments, a validation set comprising 300 news headlines was constructed. The observed agreement between the two independent annotators was 89.00%, with a Cohen’s Kappa coefficient of 0.832. This result indicates a strong level of agreement between the annotators. The 33 headlines for which the annotators initially assigned different labels were re-evaluated through consensus to establish the final human annotations. The results are presented in Table 7.

TABLE 7. Agreement Results Between VADER Labels and Final Human Annotations

Comparison	Accuracy	Macro-F1	Cohen’s Kappa
VADER–Final Human Annotation	0.8667	0.8659	0.7975

As shown in Table 7, the VADER-generated labels and the final human annotations agreed on 260 of the 300 news headlines. The agreement between VADER and the human annotations yielded an accuracy of 86.67%, a Macro-F1 score of 0.8659, and a Cohen’s Kappa coefficient of 0.7975. These results indicate a strong level of agreement between the sentiment labels automatically generated by VADER and the independent human assessments. This finding provides independent validation supporting the use of VADER-generated labels as target classes in this study.

3.2 Crisis Detection and Time Series Analysis

Weekly negative sentiment ratios and crisis candidate weeks were calculated using news headlines with available date information in the News Category Dataset V3. The sentiment analysis results obtained from the news headlines were aggregated on a weekly basis, and temporal changes in the negative sentiment ratio were examined. A four-week moving average was used to estimate the expected trend from the resulting time series, after which crisis candidate weeks were identified using a percentile-based threshold applied to positive deviations. This approach aimed to identify periods in which negative sentiment intensity exhibited a substantial increase above the expected trend.

Figure 3 illustrates the temporal variation of the weekly negative sentiment ratio obtained in this study. As shown in the figure, the negative sentiment ratio does not remain constant throughout the observation period but exhibits noticeable fluctuations across different weeks. Although the ratio generally ranges from 0.35 to 0.45, several weeks show abrupt increases, reaching values close to 0.60. Conversely, during certain periods, the ratio declines to approximately 0.25. These findings indicate that the news agenda changes considerably over time and that the intensity of negative news content increases substantially during specific weeks.

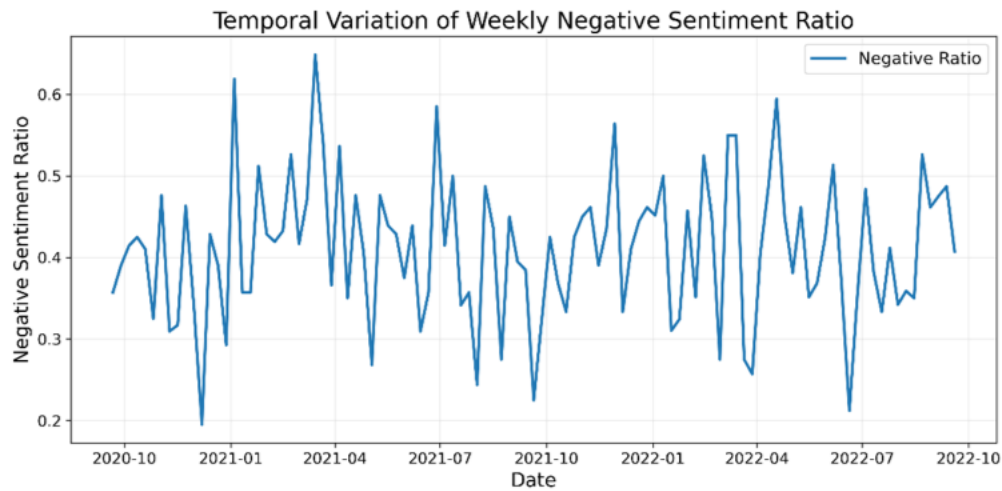


FIGURE 3. Temporal variation of the weekly negative sentiment ratio derived from news headlines

Figure 4 presents both the observed negative sentiment ratio and the expected trend estimated using the moving average. While the expected trend curve smooths short-term fluctuations and reveals the overall temporal pattern, the observed negative sentiment ratio reflects sudden changes in sentiment intensity. Periods in which the gap between the two curves becomes more pronounced indicate abnormal fluctuations that warrant further attention. In particular, weeks during which the observed negative sentiment ratio rises substantially above the expected trend suggest an unusually high concentration of negative news content. In contrast, periods in which the two curves closely follow one another represent normal variations in the news stream and reflect the system's baseline behavior.

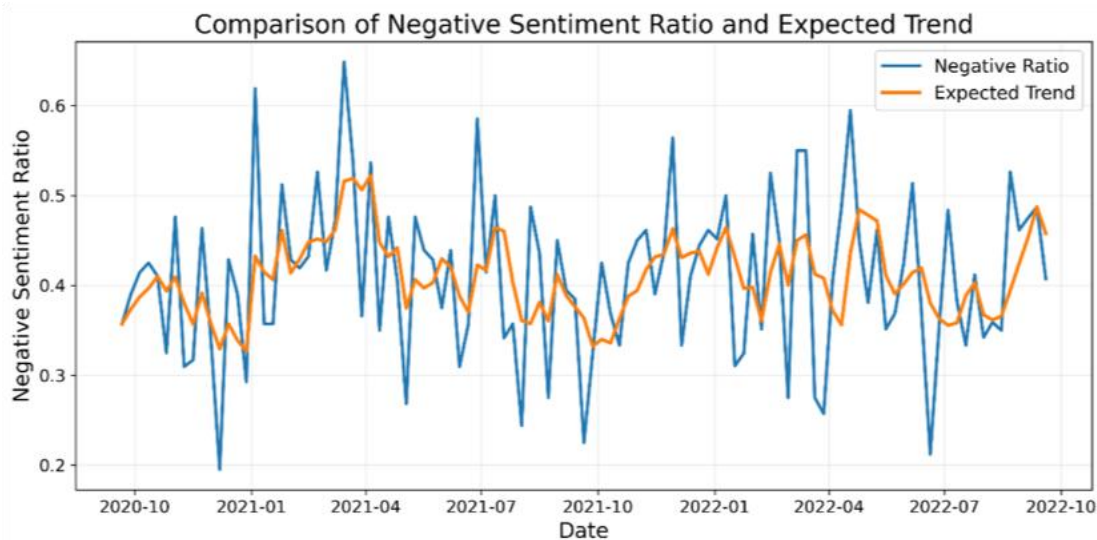


FIGURE 4. Comparison of the weekly negative sentiment ratio and the expected trend

Positive deviations between the observed negative sentiment ratio and the expected trend were calculated, and a 95th-percentile threshold was determined based exclusively on the distribution of positive deviations. Weeks whose positive deviations exceeded

this threshold were flagged as crisis candidates. This approach excluded changes below the expected trend from consideration and focused only on substantial upward increases in negative sentiment intensity.

Figure 5 illustrates the crisis candidate weeks identified by applying the 95th-percentile threshold to positive deviations of the observed negative sentiment ratio above the expected trend estimated using a four-week moving average. The analysis identified three crisis candidate weeks, beginning on January 4, 2021; June 28, 2021; and April 18, 2022. The negative sentiment ratios for these weeks were 0.6190, 0.5854, and 0.5946, respectively. Their corresponding positive deviations were 0.1864, 0.1626, and 0.1585, respectively, all exceeding the calculated threshold of 0.140434.

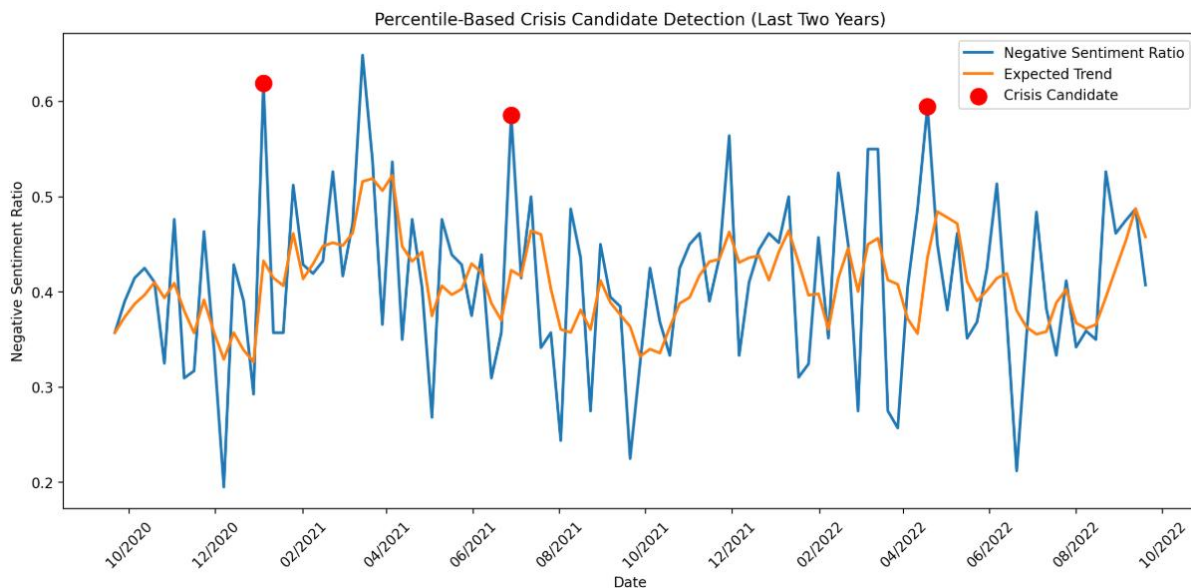


FIGURE 5. Percentile-Based Crisis Candidate Detection Based on Positive Deviations in the Negative Sentiment Ratio

A sensitivity analysis was conducted to evaluate the effects of the moving-average window length and percentile threshold on crisis candidate detection. Accordingly, different combinations of three-, four-, and five-week moving-average windows with P90, P95, and P97.5 thresholds were examined. The results are presented in Table 8.

TABLE 8. Sensitivity Analysis for Moving-Average Window Lengths and Percentile Thresholds

Window Length	Percentile Threshold	Threshold Value	Number of Crisis Candidates
3 Weeks	P90	0.124494	5
3 Weeks	P95	0.134992	3
3 Weeks	P97.5	0.160847	2
4 Weeks	P90	0.130502	6
4 Weeks	P95	0.140434	3
4 Weeks	P97.5	0.161171	2
5 Weeks	P90	0.124982	6
5 Weeks	P95	0.156220	3
5 Weeks	P97.5	0.184229	2

The sensitivity analysis results indicate that the percentile threshold plays a decisive role in determining the number of crisis candidates. Because the P90 threshold also captures relatively smaller deviations, it identified five or six weeks as crisis candidates, whereas the more stringent P97.5 threshold identified only two weeks across all window lengths. At the P95 threshold, three crisis candidates were identified for each of the three-, four-, and five-week window lengths. Furthermore, when P95 was applied with four- and five-week windows, the same weeks beginning on January 4, 2021; June 28, 2021; and April 18, 2022 were identified. These findings indicate that the results obtained with the four-week window and P95 threshold used as the baseline configuration remained largely stable under nearby parameter variations.

To assess the relationship between the identified weeks and actual news content, the negative news headlines published during these periods were examined in terms of their content. The increase observed during the week of January 4, 2021 was found to coincide with news coverage of the U.S. Capitol attack and the subsequent political violence. The increase during the week of June 28, 2021 was primarily associated with coverage of the Miami condominium collapse, along with disaster- and accident-related news reported during the same period. Similarly, the increase during the week of April 18, 2022 coincided with developments related to the Russia–Ukraine War, particularly events in eastern Ukraine and Mariupol. These correspondences indicate that the statistically detected increases were associated with periods characterized by intensified coverage of major adverse real-world events. However, this content-based examination should not be interpreted as causal validation; rather, it represents an assessment of the temporal and contextual alignment between the identified crisis candidate weeks and the prevailing news agenda. Table 9 presents the content-based evaluation of the identified crisis candidate weeks.

TABLE 9. Content-Based Evaluation of the Identified Crisis Candidate Weeks

Crisis Candidate Week	Negative Ratio	Expected Trend	Positive Deviation	Corresponding Major Event
January 4–10, 2021	0.6190	0.4326	0.1864	U.S. Capitol attack and political violence coverage
June 28–July 4, 2021	0.5854	0.4228	0.1626	Miami condominium collapse and concurrent disaster/accident coverage
April 18–24, 2022	0.5946	0.4361	0.1585	Attacks in eastern Ukraine and developments in Mariupol

4. DISCUSSION AND CONCLUSION

In this study, an integrated approach was developed to identify unusual increases in negative sentiment intensity as crisis candidates using sentiment analysis results derived from news headlines. The news headlines were first labeled as positive, neutral, or negative using VADER. Subsequently, TF-IDF and MiniLM features were extracted, and VADER-generated labels were predicted using Logistic Regression, Multinomial Naive Bayes, Linear SVM, Voting Ensemble, and the proposed hybrid Linear SVM model based on TF-IDF + MiniLM features. In the crisis candidate detection stage, weekly negative sentiment ratios calculated from the VADER labels were analyzed as a time series. The expected trend was estimated using a four-week moving average, and unusual increases in negative sentiment were identified using a percentile-based thresholding method.

The performance results of the machine learning models showed that the proposed TF-IDF + MiniLM-based hybrid Linear SVM model achieved an accuracy of 91.22% and a Macro-F1 score of 0.9099 in predicting the sentiment labels automatically generated by VADER. In stratified 5-fold cross-validation, the model achieved a mean accuracy of 0.9104 ± 0.0008 and a mean Macro-F1 score of 0.9081 ± 0.0009 . In comparison with the model using only TF-IDF features, the two-sided exact McNemar test yielded $p=0.0045$, indicating that the modest performance improvement provided by the hybrid model was statistically significant. Furthermore, validation of the VADER labels against independent human annotations yielded an accuracy of 86.67%, a Macro-F1 score of 0.8659, and a Cohen's Kappa coefficient of 0.7975. These results demonstrate a strong level of agreement between VADER-generated labels and human assessments, supporting the use of the automatically generated labels as target variables in the classification stage of the study. Nevertheless, the performance metrics of the machine learning models should be interpreted as their ability to learn and predict VADER-generated target labels rather than as a direct measure of their performance in predicting human-annotated sentiment.

The time-series analysis revealed substantial increases in the negative sentiment ratio of news headlines during specific periods. Positive deviations between the observed negative sentiment ratio and the expected trend estimated using a four-week moving average were calculated, and the 95th percentile of these deviations was used as the threshold. The analysis identified the weeks beginning on January 4, 2021; June 28, 2021; and April 18, 2022 as crisis candidates. A content-based examination of the news headlines showed that these weeks coincided with the U.S. Capitol attack, the Miami condominium collapse, and developments in the Russia–Ukraine War, respectively. Thus, the statistical increases identified by the percentile-based approach were found to coincide with periods characterized by intensified coverage of major adverse real-world events. However, these findings do

not demonstrate that the method predicts such events before they occur; rather, they indicate that the method can identify unusual increases in negative news intensity as potential crisis candidates. In this respect, the proposed approach may serve as a supporting indicator in media monitoring, crisis management, and decision-support systems.

However, this event-level examination constitutes qualitative validation rather than a quantitative evaluation against an independently annotated crisis ground truth. Since the News Category Dataset V3 does not provide exhaustive temporal crisis labels indicating crisis and non-crisis periods, crisis-level precision, recall, F1-score, and false-alarm rates could not be reliably calculated. Therefore, although the identified candidate weeks coincided with major adverse real-world events, these findings should not be interpreted as evidence of general crisis-detection accuracy. Validation against an independently annotated temporal crisis benchmark is required to quantitatively assess the detection performance of the proposed framework.

Another limitation is that the experimental comparison does not include fully fine-tuned large Transformer baselines such as BERT or RoBERTa. The present study deliberately focuses on comparatively lightweight representations and conventional classifiers, as its primary objective is to investigate whether complementary lexical and compact semantic features can support computationally efficient sentiment-based temporal monitoring rather than to maximize classification performance through large end-to-end Transformer models. Nevertheless, direct benchmarking against fine-tuned BERT- and RoBERTa-based classifiers would provide a more comprehensive assessment of the trade-off between predictive performance and computational efficiency and is therefore an important direction for future work. In addition, as a lexicon- and rule-based sentiment analysis method, VADER may have limitations in capturing context-dependent, implicit, or nuanced sentiment expressions in short news headlines. Although the human-annotation analysis provides empirical support for the reliability of the generated labels, it does not eliminate the inherent limitations of lexicon-based sentiment labeling.

The present study also does not provide a systematic computational-efficiency benchmark in terms of training time, inference latency, memory consumption, or hardware utilization across the evaluated approaches. Therefore, although MiniLM was selected as a compact semantic representation, no quantitative claim is made regarding the computational superiority of the proposed framework over larger Transformer-based alternatives. Future studies should explicitly evaluate the trade-off between predictive performance and computational cost under standardized hardware conditions.

Nevertheless, this study has several limitations. First, only news headlines were used, while full news content, social media posts, and multiple data sources were not incorporated into the analysis. Moreover, the analyses were conducted on a single dataset, and extensive external validation using different datasets was not performed. This limitation should be considered when assessing the generalizability of the findings across different data sources.

Another limitation is the use of sentiment labels automatically generated by VADER as target labels for training the classification models. Although strong agreement between VADER and human assessments was observed in the human-annotated validation set of 300 news headlines, the relatively limited size of this validation set compared with the main dataset should be acknowledged. Future studies should employ larger datasets annotated by domain experts and evaluate model performance directly against human-annotated sentiment labels.

Finally, because VADER relies on an English-language sentiment lexicon, the present analysis was limited to English news data. Future studies may extend the proposed framework to other languages by employing language-specific sentiment analysis tools or validated multilingual models.

Future research will focus on extending the proposed framework in several directions. First, multilingual crisis-candidate analysis can be investigated by incorporating language-specific sentiment resources or validated multilingual language models, enabling the framework to monitor news streams across different languages and geographical contexts. Second, the temporal anomaly-detection mechanism can be integrated with external knowledge graphs or structured event representations to move beyond general crisis-candidate identification toward contextual interpretation and crisis-type classification. Third, future studies should evaluate the framework in real-time news-streaming environments, with particular attention to inference latency, throughput, memory requirements, and the trade-off between computational cost and detection responsiveness. The integration of real-time news feeds and complementary sources such as social media may further improve the temporal coverage of the framework. Finally, a human-in-the-loop validation mechanism can be incorporated in which automatically identified crisis candidates are reviewed by domain experts, allowing contextual verification of detected signals and providing feedback for subsequent refinement of the detection mechanism. Future evaluations should also include independent datasets representing different crisis domains and, where available, temporally annotated crisis benchmarks.

In conclusion, the approach proposed in this study combines sentiment analysis results derived from news headlines with time-series analysis and a percentile-based thresholding method to identify unusual increases in negative sentiment intensity. The correspondence between the three crisis candidate weeks identified based on positive deviations and major adverse real-world events indicates that the statistically detected increases coincide with meaningful periods in the news agenda. The proposed approach should not be regarded as a direct crisis prediction system, but rather as a practical decision-support mechanism designed to facilitate the monitoring of potential crisis periods and support their subsequent evaluation by experts.

Ethics Statement

This study used the publicly available News Category Dataset V3 and did not involve experiments on human or animal subjects. The human annotation task was limited to labelling the sentiment of publicly available news headlines; no personal or identifiable data were collected from the annotators. Therefore, ethical approval was not required for this study.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Sena Nur Çulcu: Conceptualization, Methodology, Software, Data curation, Formal analysis, Investigation, Visualization, Writing—original draft. Canan Koç: Conceptualization, Methodology, Validation, Supervision, Writing—review & editing. All authors have read and agreed to the final version of the manuscript.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Data Availability Statement

The data used in this study are publicly available. The News Category Dataset V3 can be obtained from publicly accessible repositories. The processed data and source code used in this study are available from the corresponding author upon reasonable request.

Acknowledgments

The authors acknowledge the use of generative artificial intelligence (AI) tools solely for language editing and English translation of the manuscript. All scientific content, methodology, experimental design, analysis, and conclusions were developed and verified by the authors.

REFERENCES

1. Hamza, A., Muhammad, A., Abbas, M. S., & Jan, S. U. (2026). Contextual deep learning for accurate news article categorization with pre-trained embeddings. *Scientific Reports*, 16(1). <https://doi.org/10.1038/s41598-026-38998-3>
2. Zare, M. (2025). *Attention-guided feature fusion (AGFF) model for integrating statistical and semantic features in news text classification* (arXiv:2511.17184). arXiv. <https://doi.org/10.48550/arXiv.2511.17184>
3. Aprilio, P., Felix, M., Nugraha, P. S., & Fahmi, H. (2025). Hybrid feature combination of TF-IDF and BERT for enhanced information retrieval accuracy. *JISA (Jurnal Informatika dan Sains)*, 8(1), 8–15. <https://doi.org/10.31326/jisa.v8i1.2179>
4. Zhang, G., & Hu, J. (2024). Enhanced industrial text classification via hyper variational graph-guided global context integration. *PeerJ Computer Science*, 10, Article e1788. <https://doi.org/10.7717/peerj-cs.1788>
5. Chhetri, G., Das, S., & Chowdhury, T. I. (2025). *WISE: Web information satire and fakeness evaluation* (arXiv:2512.24000). arXiv. <https://doi.org/10.48550/arXiv.2512.24000>
6. Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2024). Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3881–3906). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.246>
7. Kuila, A., & Sarkar, S. (2024). Deciphering political entity sentiment in news with large language models: Zero-shot and few-shot strategies. In *Proceedings of the 1st Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2024)*.
8. Dharrao, D., MR, A., Mital, R., Vengali, A., Pangavhane, M., Rajput, S., & Bongale, A. M. (2024). An efficient method for disaster tweets classification using gradient-based optimized convolutional neural networks with BERT embeddings. *MethodsX*, 13, Article 102843. <https://doi.org/10.1016/j.mex.2024.102843>
9. Yicheng, S., Louw, S. A., Thant, Y. M., Arya, A., Konduru, R. T., & Avtar, R. (2026). BART-Disaster-X: An effective model to extract information from disaster-related tweets on X. *Remote Sensing Applications: Society and Environment*, Article 102174. <https://doi.org/10.1016/j.rsase.2026.102174>
10. Jahan, I., Islam, M. N., Hasan, M. M., & Siddiky, M. R. (2024). Comparative analysis of machine learning algorithms for sentiment classification in social media text. *World Journal of Advanced Research and Reviews*, 23(3), 2842–2852. <https://doi.org/10.30574/wjarr.2024.23.3.2983>
11. Peng, L.-J., Shao, X.-G., & Huang, W.-M. (2021). Research on the early-warning model of network public opinion of major emergencies. *IEEE Access*, 9, 44162–44172. <https://doi.org/10.1109/ACCESS.2021.3066242>
12. Dui, Y., & Hu, H. (2024). Social media public opinion detection using multimodal natural language processing and attention mechanisms. *IET Information Security*, 2024(1). <https://doi.org/10.1049/2024/8880804>
13. Kahraman, İ., & Önder, E. (2026). A crisis that began with a tweet: Temporal and thematic analysis of social media sentiment dynamics in financial collapse processes. *Yalova Sosyal Bilimler Dergisi*, 16(1), 88–110. <https://doi.org/10.17828/yalovasosbil.1785839>

14. Kırıcı, P., & Gülbak, E. (2020). Instagram verileri ile duygu analizi [Sentiment analysis with Instagram data]. *Avrupa Bilim ve Teknoloji Dergisi (EJOSAT)*, 360–364. <https://doi.org/10.31590/ejosat.780129>
15. Çılgin, C., Baş, M., Bilgehan, H., & Ünal, C. (2022). Twitter sentiment analysis during Covid-19 outbreak with VADER. *AJIT-e: Academic Journal of Information Technology*, 13(49), 72–89. <https://doi.org/10.5824/ajite.2022.02.001.x>
16. Kınık, D., & Güran, A. (2021). TF-IDF ve Doc2Vec tabanlı Türkçe metin sınıflandırma sisteminin başarımlarının ardışık kelime grubu tespiti ile artırılması [Improving the performance of a TF-IDF and Doc2Vec-based Turkish text classification system through consecutive word group detection]. *Avrupa Bilim ve Teknoloji Dergisi (EJOSAT)*, (21), 323–332. <https://doi.org/10.31590/ejosat.774144>
17. Çelik, Ö., & Koç, B. C. (2021). TF-IDF, Word2vec ve Fasttext vektör model yöntemleri ile Türkçe haber metinlerinin sınıflandırılması [Classification of Turkish news texts using TF-IDF, Word2vec and FastText vector model methods]. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 23(67), 121–127. <https://doi.org/10.21205/deufmd.2021236710>
18. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4656–4667). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.372>
19. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
20. Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, 3, 238–248. <https://doi.org/10.1016/j.susoc.2022.03.001>
21. Yang, J. (2026). Methods of improving sentiment analysis based on sentiment dictionary. *Franklin Open*, 15, Article 100616. <https://doi.org/10.1016/j.fraope.2026.100616>
22. Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621–633. <https://doi.org/10.1016/j.eswa.2012.07.059>
23. Aburomman, A. A., & Reaz, M. B. I. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. *Applied Soft Computing*, 38, 360–372. <https://doi.org/10.1016/j.asoc.2015.10.011>
24. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>