information

MDPI

Article

A Comparative Framework for Political Violence Event Classification Using Machine Learning, Deep Learning, and Zero-Shot Language Models

Ujala Beenish [1], Saadia Ishtiaq Nauman [2], Sadaf Abdul Rauf [1,3,*], Fatima Mumtaz [1], Muhammad Ghulam Abbas Malik [4], Muhammad Imran [5,6] and Muddesar Iqbal [7]

[1] Speech and Language Processing Group, Department of Computer Science, (FJWU-IIUI), Islamabad 44000, Pakistan
[2] Department of Communication and Media Studies, Fatima Jinnah Women University, Rawalpindi 46000, Pakistan
[3] Department of Computer Science, International Islamic University, Islamabad 44000, Pakistan
[4] Interdisciplinary Sustainable Systems (IS2) Group, College of Computer and Information Sciences, Prince Sultan University, Riyadh 12435, Saudi Arabia; amaalik@psu.edu.sa
[5] Educational Research Lab, Prince Sultan University, Riyadh 12345, Saudi Arabia; mimran@psu.edu.sa
[6] Department of English, Khazar University, Baku AZ1096, Azerbaijan
[7] Smart Systems Lab, College of Engineering, Prince Sultan University, Riyadh 12435, Saudi Arabia
* Correspondence: sadaf.abdulrauf@gmail.com

Abstract

Political violence poses a significant challenge to global stability, underscoring the need for comparative analytical models that support analytical interpretation of structured conflict data. This paper presents a comparative evaluation of 12 machine learning approaches, including traditional supervised models, deep learning architectures, and zero-shot large language models, for the classification of political violence events using the Armed Conflict Location and Event Data Project (ACLED) dataset (2010–2020, over 40,000 events). The results demonstrate that, on short structured event text represented via TF-IDF, fine-tuned traditional machine learning models achieve stronger performance than zero-shot LLM approaches and deep learning models on structured event data. We further introduce a multilingual classification framework for English and Urdu news content, illustrating cross-lingual transfer robustness using machine-translated Urdu data; results reflect translation-based evaluation conditions and should not be interpreted as performance on naturally occurring low-resource Urdu political-event text. As an exploratory extension, the framework is applied to 57,700 tweets related to the Article 370 crisis in Kashmir to illustrate applicability to unstructured social media text; given that the best Twitter model (55% accuracy) falls below the 69% majority-class baseline, these results should be interpreted solely as coarse discourse indicators and not as a validated classification component. Unlike prior work, this study systematically combines multilingual evaluation with zero-shot LLM analysis for political event classification. Geographic out-of-sample validation (leave-one-country-out or leave-one-region-out) was not conducted; the reported performance should therefore not be interpreted as evidence of cross-regional generalizability without further experimentation. The findings highlight practical considerations for designing data-driven analytical frameworks for conflict monitoring and analytical decision support.

Keywords: political violence classification; machine learning; large language models (LLMs); multilingual classification; deep learning

Check for updates

Academic Editor: Omar El-Gayar

Received: 2 April 2026
Revised: 30 May 2026
Accepted: 5 June 2026
Published: 23 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Political violence has escalated globally, with devastating consequences for communities and governance structures [1]. Countries experiencing civil wars and international disputes face heightened risks of political violence, causing sudden damage to civilians, infrastructure, and institutional stability [2,3]. The increasing trend of violent outbursts has underscored the need for machine learning approaches capable of accurately detecting and classifying such events [4,5]. Despite growing interest in transformer-based and large language models, their effectiveness for structured political event classification and multilingual low-resource settings remains insufficiently evaluated.

Studies [6,7] demonstrate that news events act as early warning indicators of wars among and within states. Using machine learning modeling techniques, it is possible to capture the complexities and dynamics of conflict in real time and accurately identify trends, risks, and root causes with greater analytical precision [8,9]. Recent advancements in machine learning and the availability of widespread news event datasets have made it possible to tackle this problem [10,11]. These datasets, often assembled from very large collections of press reports, are valuable for offering a real-time window into the dynamics of civil unrest, terrorism and other forms of politically motivated conflict [12].

Large-scale news datasets such as GDELT [13] and ICEWS [14] have been widely used for early conflict prediction. For example, this study [13] presents the Global Database of Events, Language, and Tone (GDELT), which has been widely used to investigate cases such as how conflict escalation and global unrest unfold, unlike another study [14] that focuses on the ICEWS dataset, which concentrates on political crises and protests for the development of early warning systems for the prediction of conflict.

While both datasets have primarily supported conflict forecasting and event extraction, the present study addresses a distinct task: event classification, which involves the assignment of already-coded, structured ACLED event records to predefined ontological categories. This distinction is important because classification operates on verified, observed data, whereas forecasting involves temporal prediction under uncertainty. Additionally, the Twitter-based component of the present study addresses stance detection and sentiment analysis, which involves classifying subjective discourse orientation in user-generated social media content. This differs fundamentally from structured event classification in terms of data characteristics, noise levels, annotation complexity, and expected performance.

In this study, we used the Armed Conflict Location and Event Data Project (ACLED) derived from news reports [12]. Many types of incidents are included in these events, ranging from protests, riots, and acts of terrorism [15,16]. The volume and diversity of these incidents motivates the development of systematic classification methods. ACLED's regional specificity and fine-grained annotations make it a suitable dataset for evaluating machine learning and multilingual classification approaches.

Classifying and analyzing political violence requires accounting for complex political factors such as ethnic tensions, socio-economic disparities, and historical grievances [17]. Strategies adopted by both political parties and different interest groups to accomplish their objectives usually produce unintended escalations in violence [18,19]. However, the effectiveness of different modeling paradigms for structured political event data remains empirically underexplored.

The advent of LLMs has added the capabilities of zero-shot classification; models can be used to classify text without any task-specific training [20]. Although these models provide convenience in deployment, their effectiveness compared to the fine-tuned traditional methods of domain-specific tasks is not empirically substantiated.

This gap is especially pertinent to the classification of political violence, where precision has a direct influence on analytical interpretation of structured conflict data. Under-

standing whether general-purpose zero-shot models are as capable as task-specific fine-tuned models has important theoretical implications for real-world deployment.

Throughout this study, 'classification' refers to the assignment of already-coded ACLED event records to predefined categories (protests, battles, riots, violence against civilians, explosions/remote violence, and other); it does not constitute temporal forecasting of future, unobserved events. The temporal split (training on 2010–2019, evaluating on 2020) reflects a realistic out-of-sample evaluation condition rather than a prospective forecasting design.

We used machine learning and large-scale datasets to evaluate approaches for political violence classification. The framework is evaluated through systematic comparison of traditional supervised models, deep learning architectures, and zero-shot LLMs. Traditional methods achieved higher accuracy compared to deep learning models and LLMs. However, the BERT–Multilingual framework performed effectively on Urdu and English datasets, demonstrating that it can analyze low-resource languages [4,21,22]. These findings contribute to the comparative understanding of political violence event classification and have implications for the design of classification frameworks in conflict research contexts [23,24].

The problem of classifying political violence events using machine learning, deep learning, transformer-based models, and zero-shot LLMs has been addressed. However, existing studies rarely test multilingual or zero-shot LLMs for structured political event data. This study bridges that gap by systematically comparing traditional ML, deep learning, and LLM approaches on the ACLED dataset. The main contributions are as follows:

- It systematically compares traditional machine learning, deep learning, and zero-shot LLMs for political violence classification using identical temporal splits on ACLED data (2010–2020).
- It evaluates multilingual performance on English and Urdu, demonstrating that XLM-RoBERTa substantially outperforms Multilingual BERT on Urdu due to differences in pretraining data scale and subword modeling.
- The proposed framework is applied as an illustrative extension to Twitter data from the Article 370 crisis, providing exploratory discourse indicators; given that the best model (55% accuracy) falls below the 69% majority-class baseline, this component does not constitute a validated classification finding.

Building on these distinctions, the present study addresses three explicit research questions:

1. For structured political event classification using ACLED data, do traditional supervised models outperform transformer-based and zero-shot LLM approaches under temporal out-of-sample evaluation?
2. Does multilingual transfer learning enable robust political event classification in a low-resource language setting, and which architectural choices determine cross-lingual performance?
3. Does social media stance classification provide complementary discourse analysis signals relative to structured event data, and what are its practical limitations for operational deployment?

The rest of this paper is organized as follows: The proposed political violence classification framework is introduced in Section 2. The comparison of zero-shot and fine-tuned models, data sources, preprocessing techniques and the feature extraction process are described in Section 3. The experimental setup is presented in Section 4. The results for the political violence classification task are presented in Section 5. Section 6 presents an exploratory extension applying the framework to Twitter data from the Article 370 crisis to

illustrate applicability to unstructured social media text. The paper concludes with discussion in Sections 7 and 8 stating key findings and future directions.

2. Political Violence Classification Framework

The political violence classification framework is developed to systematically classify political violence events. The structure follows a pipeline of data preprocessing, feature extraction and machine learning-based classification. The framework handles diverse event types including protests, battles, and violence against civilians.

The process begins with data cleaning, which removes redundant entries and handles missing values to prepare data for analysis. The classification target is event_type. Key textual features are extracted from the notes field via TF-IDF. Event type, sub-event type, location, and fatality rate are used for dataset characterization only and are not included as model inputs. Sub-events (e.g., arrests, bombings and suicide attacks) were grouped under broader event types such as protests or violence. Geographic areas of political violence are identified on the basis of location data; fatality rates convey the severity of the events. Each event was contextually verified with news sources.

TF-IDF features were extracted from textual fields and used as input to all traditional classifiers. These features were then used to train different classification models for classifying political violence events. Initially, the traditional machine learning models such as Random Forest, Logistic Regression, and Naïve Bayes, are evaluated. These models can act as a baseline for a classification task and help to understand the most important patterns in the data.

However, to represent the complexity of political violence and identify the relationships behind the data, we use Long Short-Term Memory (LSTM) networks and transformer-based models such as BERT, Roberta and Multilingual BERT. Specifically, these models are best suited for multilingual, sequential, and rich-context text data. Figure 1 illustrates the overall framework for classifying political violence events underlying this study. The framework covers the whole pipeline, from data collection and preprocessing to feature engineering and model building using both traditional and deep learning techniques.

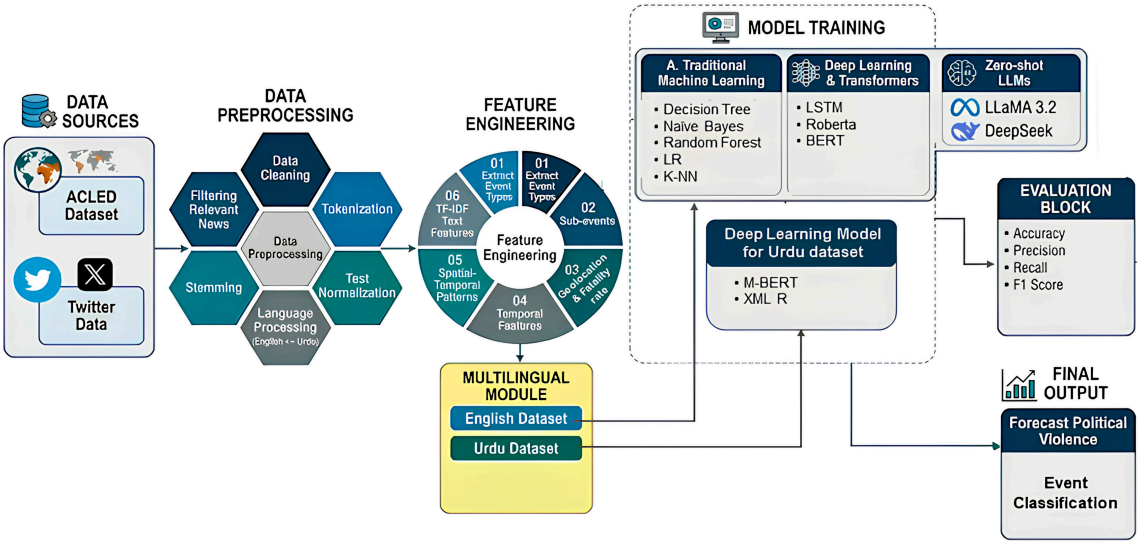


Figure 1. Architecture of the framework proposed for political violence event classification. It includes data processing (cleaning, tokenization, lemmatization, stemming, etc.), feature engineering (TF-IDF extraction from event notes field; event_type and sub_event_type used as label and dataset description only, not as model inputs), and model training of classification models. By identifying categorical patterns in structured event data, this process is intended to generate insights and support accurate classification of political violence event types.

3. Comparison of Zero-Shot LLMS and Fine-Tuned Models

We compared two methods to assess domain-specific training and general-purpose inference: (1) train traditional models and deep learning models on ACLED data (Logistic Regression, Random Forest, SVM, BERT, RoBERTa), and (2) zero-shot LLMs used without task-specific training (DeepSeek, LLaMA 3.2 3B Instruct). This comparison evaluates the effectiveness of domain-specific fine-tuning versus general-purpose pre-trained models in political violence classification.

3.1. Data Preprocessing and Feature Extraction

In this study, the Armed Conflict Location Event Data Project (ACLED) was used, which spans a decade (2010–2020) and contains more than 40,000 political events [13].

The dataset includes fine-grained annotations covering event type, geographic location, dates, and fatality rate, making it well suited for training classification models [25]. These features help to identify patterns using machine learning frameworks that can accurately classify political violence event types.

ACLED gathers data from a wide range of sources which include national and subnational media, international agencies, and regional reports and local partners. The events covered through these sources are on different scales and origins; hence, diversity is guaranteed. Table 1 contains a breakdown of each of these source categories and the number of events reported at each.

Table 2 is a summary of the event categories extracted from the dataset. From this, we get a clear outline of types and frequency of political violence events. Such insights are crucial for understanding dynamics of political violence and, by implication, which event categories to prioritize for classification modeling. ACLED provides structured, event-level data including event_date, location, event_type, sub_event_type, fatalities and notes.

It does not include full-text news articles or tweets; instead, it codes events based on open-source news reports. The dataset was downloaded from the official ACLED website https://acleddata.com/data/ (accessed on 1 January 2022). The preprocessing pipeline is shown in Algorithm 1. In preprocessing steps, it removed duplicate records, filtered text content (such as emojis and URLs), and excluded irrelevant languages (such as Hindi). These steps were taken to ensure that the final dataset contained only relevant political discourse. The text normalization (lowercasing, stemming, and lemmatization) was applied to reduce word variants (e.g., ‘run’, ‘running’, ‘ran’) to root forms.

For consistency, all text was changed to lowercase, and stopwords (common words that offer little analytical value, e.g., ’and’, ’the’, etc.) were removed. Stemming and lemmatization were used to reduce words to their root forms, and treatments of different forms of the same word (e.g., run, running, and ran) were also made as a single entity.

Standardization of the linguistic structure increased the accuracy and reliability of subsequent analyses. The features were critical for pulling out the sentiment, political leanings, and event-specific information inherent in the dataset.

Algorithm 1: Data Preprocessing and Feature Extraction

```
Input: ACLED_data, twitter_data
Output: clean_acled_data, tfidf_acled, tfidf_twitter, processed_tweets
# ACLED Preprocessing
Function preprocess_acled(acled_data):
    filtered_data = filter_features(acled_data,
                        ["event_type", "sub_event_type",
                          "location", "fatality_rate", "event_date"])
    cleaned_data = handle_missing_values(filtered_data)
    cleaned_data = remove_duplicates(cleaned_data)
    cleaned_data = exclude_irrelevant_languages(cleaned_data)
    text_corpus = extract_text_fields(cleaned_data)
    text_corpus = to_lowercase(text_corpus)
    text_corpus = remove_stopwords(text_corpus)
    text_corpus = lemmatize(text_corpus)
    text_corpus = stem(text_corpus)
    tfidf_acled = TfidfVectorizer().fit_transform(text_corpus)
    return cleaned_data, tfidf_acled
# Twitter Preprocessing
Function preprocess_twitter(twitter_data):
    processed_tweets = []
    For each tweet in twitter_data:
        tweet_text = extract_text(tweet)
        source = extract_source(tweet)
        location = extract_location(tweet)
        hashtags = extract_hashtags(tweet)
        tweet_text = remove_links(tweet_text)
        tweet_text = to_lowercase(tweet_text)
        tweet_text = remove_symbols(tweet_text)
        tweet_text = remove_stopwords(tweet_text)
        tweet_text = lemmatize(tweet_text)
        processed_tweets.append(
            {'tweet_text': tweet_text, 'source': source, 'location': location, 'hashtags':
hashtags})
    tfidf_twitter = TfidfVectorizer().fit_transform([t.tweet_text for t in
processed_tweets])
    return tfidf_twitter, processed_tweets
# Main
clean_acled_data, tfidf_acled = preprocess_acled(ACLED_data)
tfidf_twitter, processed_tweets = preprocess_twitter(twitter_data)
```

3.2. Urdu Text Preprocessing

To preprocess Urdu text, we applied standard cleaning techniques: stopword removal, punctuation elimination, and handling of non-standard characters. To enable cross-lingual modeling, the complete ACLED dataset (approximately 58,000 records spanning 2010–2020) was translated from English to Urdu using Google Translate API, maintaining identical temporal splits as the English dataset: 50,000 training records (2010–2019) and 8000 test records (2020). To validate translation quality, 500 stratified samples (0.86% of the corpus) were manually reviewed by two independent bilingual annotators with expertise

in Computational Linguistics and South Asian Political Studies, achieving inter-annotator agreement of $\kappa = 0.82$. Each translation was assessed on semantic accuracy, contextual appropriateness, and grammatical correctness.

Common translation errors included political terminology inconsistency, where "militants" was variably translated as "جنگجو" (fighters) versus "عسکریت پسند" (militants); contextual ambiguity in legal terms, where "Section 144" was directly transliterated as "144 سیکشن " without explanatory context; and named entity transliteration variation, with place names like "Srinagar" inconsistently rendered as "سری نگر" versus "شری نگر ". Manual review found 90.6% of translations acceptable without modification, while 9.4% required minor terminology standardization. Identified inconsistencies were systematically corrected across the full dataset using pattern matching and manual verification, ensuring that both M-BERT and XLM-RoBERTa were trained and evaluated on standardized, high-quality Urdu text. This structure of the data allowed us to extract meaningful insights and identify trends, biases, and underlying sentiment within the political discourse, offering a comprehensive understanding of public discourse dynamics.

Table 1. Distribution of news sources by scale in ACLED dataset.

No.	Source Scale	Values
1.	National	37,407
2.	Subnational	19,281
3.	Other	243
4.	International	219
5.	Subnational-National	176
6.	Regional	153
7.	Local partner-Other	92
8.	National-Regional	70
9.	National-International	54
10.	Other-National	39
11.	Other-Subnational	13
12.	New media	10
13.	Regional-International	5
14.	Subnational-International	3
15.	Other-International	3
16.	Other-Regional	2
17.	Subnational-Regional	1

It should be noted that the Urdu corpus employed in this study is a machine-translated version of the English ACLED dataset rather than an independently collected Urdu political-event corpus. Consequently, the results reported for Urdu reflect cross-lingual transfer robustness under controlled translation conditions and should not be generalized to performance on naturally occurring Urdu political-event text, which would differ substantially in vocabulary distribution, syntactic register, and named entity characteristics.

The target variable for all classification experiments is event_type, encompassing six predefined categories: protests, battles, riots, violence against civilians, explosions/remote violence, and other. The sole input to all traditional machine learning and deep learning models consists of TF-IDF representations derived from the ACLED notes field, which contains free-text event descriptions sourced from news reports. Structured metadata fields such as sub_event_type, location, fatality_rate, and event_date were used exclusively for dataset characterization and exploratory analysis; they were explicitly excluded from the predictor set to prevent label leakage. For the Twitter-based case study, features including tweet_text, source, hashtags, and location were extracted and processed via TF-IDF. All TF-IDF vectorization was performed using default scikit-learn settings: max_features set

to None to retain the full vocabulary, ngram_range of (1,1) to capture unigrams, min_df of 1, and max_df of 1.0. A random seed of 42 was set across all experiments to ensure full reproducibility of results. This seed was applied consistently across all experiments including transformer fine-tuning (PyTorch manual seed (version 2.0.1), NumPy (version 1.24.3) seed, and Python (version 3.10.12) random seed) to ensure full reproducibility.

Table 2. Number of different political violence events that occurred between (2010 and 2020).

Event	Count
Protests	42,726
Explosions/Remote violence	4535
Battles	4115
Violence against civilians	3123
Riots	2593
Other	678

4. Experiment Setup

The experiments were conducted using the ACLED dataset that contains information on political violence events from 2010 to 2020. The dataset has entries classified into categories such as protests, battles, violence against civilians, and additional sub-event types, for example, arrests, drone strikes, and shelling. The dataset also includes information such as event type, the number of fatalities, locations, and date. Both English and Urdu datasets were used for the experiments. For cross-lingual evaluation, the English dataset was translated into Urdu using Google Translator API.

The ACLED dataset from 2010 to 2019 used for training and testing was performed on 2020 data to determine their generalization performance on unseen events. To make the evaluation more robust, a temporal split strategy was employed. Models were trained on data from 2010 to 2019 and evaluated on 2020 data. Temporal splitting was preferred over k-fold cross-validation, as random shuffling of conflict events would allow future events to leak into the training set, producing overly optimistic performance estimates. This represents a conservative out-of-sample evaluation design.

Geographic out-of-sample validation was not conducted in the present study. Although the ACLED dataset spans multiple countries and regions, implementing leave-one-country-out or leave-one-region-out evaluation strategies falls beyond the scope of the current work. We nevertheless acknowledge this as an important limitation, since political violence discourse may contain country-specific lexical and contextual regularities, including actor names and location-dependent terminology, which could artificially inflate apparent cross-country generalization performance. Future research should therefore investigate whether models trained on one region can effectively generalize to geographically unseen regions, particularly under deployment conditions involving new political contexts. Accordingly, the present findings should not be interpreted as evidence of robust cross-geographic generalization without additional region-level validation experiments.

Given the pronounced class imbalance in the dataset, macro-F1 and balanced accuracy are the primary evaluation metrics for all model comparisons in this study. Macro-F1 treats all classes equally and is sensitive to minority-class performance; balanced accuracy provides a class-imbalance-corrected alternative to overall accuracy. Weighted F1-score is reported as a supplementary aggregate indicator, and overall accuracy is included for completeness only. Per-class recall is reported in Table 3b to surface model behavior on high-consequence minority categories.

Per-class recall for each event category is reported in Table 3b to surface model behavior on high-consequence minority categories. A majority-class baseline predicting

only 'Protests' for every instance achieves approximately 74% overall accuracy but only 16.7% balanced accuracy and a macro-F1 of approximately 0.13, establishing the minimum threshold any meaningful classifier must exceed.

No resampling or balancing technique such as SMOTE was applied to the dataset, as the class distribution in ACLED reflects the natural real-world distribution of political violence events, where protests are genuinely the most frequent occurrence. Furthermore, consistency with prior ACLED-based studies [8,13,25] was maintained to ensure benchmark comparability. The effect of class imbalance on evaluation was instead addressed through weighted averaging of precision, recall, and F1-score, which assigns proportional importance to each class based on its support.

4.1. Models

In this study, several traditional models, deep learning models, LLMs and transformer-based models have been implemented and evaluated for classifying political violence events.

4.1.1. Large Language Models (LLMs)

To evaluate zero-shot classification, two LLMs representing distinct paradigms were selected. DeepSeek [26] was chosen as a closed-source, API-based frontier model accessed via its official API (api.deepseek.com, model: deepseek-chat, accessed on 1 June 2025) using default inference settings, enabling evaluation of the upper bound of commercially deployed general-purpose LLMs without task-specific training.

LLaMA 3.2 3B Instruct [27] was selected as an open-source, instruction-tuned alternative (model ID: meta-llama/Llama-3.2-3B-Instruct), deployable on standard academic hardware and ensuring full reproducibility for researchers without API access. The 3B parameter scale was deliberately chosen to represent a computationally feasible open-source model. Models such as GPT-4 were excluded due to cost constraints and limited reproducibility in academic settings. Together, these two models enable a controlled comparison between closed-source API-based and open-source locally deployed zero-shot approaches. Since the task required selecting one label from six fixed categories using a structured prompt, output variability for DeepSeek was negligible. LLaMA 3.2 3B Instruct inference was performed on Google Colab using a TPU v2 accelerator via the Hugging Face Transformers library, with generation parameters set to temperature = 0.0 (greedy decoding), max_new_tokens = 20, and do_sample = False to ensure deterministic, reproducible outputs. Standardized prompts as shown in Figure 2 were used for each zero-shot model to ensure fair evaluation. Each model output was post-processed by extracting the first matching category label from the six predefined classes using exact string matching. All outputs from both DeepSeek and LLaMA 3.2 3B Instruct were successfully mapped to a valid category; no invalid or ambiguous responses were observed across the entire test set.

4.1.2. Deep Learning Models

The Long Short-Term Memory (LSTM) model, used to capture sequential dependencies in the dataset, was also evaluated [28]. The LSTM was implemented using TensorFlow/Keras and trained on Google Colab with a TPU v2 accelerator. The architecture comprised two stacked LSTM layers with 128 hidden units each, followed by a dense softmax output layer with 6 units corresponding to the six event categories. A dropout rate of 0.3 was applied after each LSTM layer to prevent overfitting. The model was trained for 10 epochs with a batch size of 32, using the Adam optimizer (learning rate = 0.001) and categorical cross-entropy loss function. The maximum sequence length was set to 512 tokens, selected based on the 95th percentile of token counts in the dataset, and the embedding dimension was set to 64 to balance representational capacity with computational efficiency given the relatively compact ACLED vocabulary. As traditional models struggle to capture

long-term dependencies from text, this model was expected to outperform them. LSTM was specifically selected to evaluate whether sequential dependency modeling could capture temporal patterns in political event text beyond bag-of-words representations [28].

```
Analyze this event and classify it into exactly one category:
Categories:

   • Protests

   • Explosions/Remote violence

   • Battles

   • Violence against civilians

   • Riots

   • Other

Event Details:
Date: {row.get('event_date', 'N/A')}
Actors: {row.get('inter1', 'N/A')} vs {row.get('inter2', 'N/A')}
Location: {row.get('admin1', 'N/A')}
Fatalities: {row.get('fatalities', 'N/A')}
Description: {str(row.get('notes_clean', 'N/A'))[:300]}
Classification:"""
```

Figure 2. Prompt for event classification.

4.1.3. Traditional Supervised Learning Models

Initially, traditional supervised learning models were used, including Support Vector Machines (SVM) [29], Random Forest (RF) [30], Naïve Bayes (NB) [31], Decision Tree (DT) [32], Logistic Regression (LR) [33] and K-Nearest Neighbors (K-NN). To extract meaningful features from the textual data, the TF-IDF (Term Frequency-Inverse Document Frequency) technique, which converts text data to numerical feature vectors, was used to train these models. TF-IDF proved useful in identifying key patterns in political events as it measures the importance of words based upon their frequency in documents. TF-IDF was selected over dense word embedding representations such as Word2Vec or GloVe because ACLED event notes are short, factual, and domain-specific texts, making sparse lexical representations more suitable for traditional classifiers operating in high dimensional feature spaces [21]. Moreover, TF-IDF offers greater interpretability by highlighting discriminative domain-specific terms without requiring large external corpora for embedding training, which aligns with the transparency requirements of conflict monitoring applications [34].

It is important to acknowledge that this design choice introduces a structural advantage for lexical models. ACLED notes are brief, factual summaries typically one to three sentences rather than long narrative documents. This format naturally aligns with TF-IDF representations. Transformer-based models and LLMs are designed to leverage rich contextual information across longer text; when applied to short formulaic descriptions, their contextual capacity is underutilized. Therefore, the performance gap observed here reflects data-model alignment rather than a general superiority of traditional approaches.

For traditional machine learning models (SVM, Logistic Regression, Random Forest, Naïve Bayes, Decision Tree, and K-NN), we employed the default hyperparameter. Specifically, Logistic Regression used the lbfgs solver with C = 1.0 and max_iter = 1000; SVM used

an RBF kernel with C = 1.0 and gamma = 'scale'; Random Forest used n_estimators = 100 with max_features = 'sqrt'; Naïve Bayes (MultinomialNB) used default alpha = 1.0 (Laplace smoothing); Decision Tree used the gini criterion with no depth constraint; and K-NN used k = 5 with Euclidean distance. These defaults were selected rather than custom-tuned values for two reasons: (1) they represent theoretically grounded baselines, and (2) the primary objective of this study was a comparative evaluation across model families rather than optimization of individual models. Given that traditional models already achieved 96–98% accuracy with these settings. Traditional machine learning models were chosen for their simplicity, interpretability, and computational efficiency with structured data, which are essential for tasks requiring transparent decision-making [34].

4.1.4. Transformer-Based Models

BERT [35] and RoBERTa [36] were chosen as the dominant transformer-based models for English NLP tasks, enabling evaluation of deep contextual representations on the same dataset. For multilingual evaluation on Urdu, M-BERT and XLM-RoBERTa were selected because they represent two architecturally distinct cross-lingual approaches, where M-BERT relies on Wikipedia-based multilingual pretraining while XLM-RoBERTa uses a significantly larger CommonCrawl corpus, allowing direct comparison of their low-resource language capabilities [37]. DeepSeek and LLaMA 3.2 3B Instruct were included to represent zero-shot LLMs, one closed-source API-based and one open-source locally deployable, enabling a controlled comparison of general-purpose inference versus task-specific fine-tuning. In our case, the relatively stronger performance of traditional models may be attributed to the structural nature of the data and class imbalance, which are more sensitive to deep learning models.

To ensure fair performance comparison, we performed systematic hyperparameter tuning across all models. For deep learning and transformer-based architectures, a batch size of 32 and a learning rate of 2×10^{-5} were employed. A maximum sequence length of 128 tokens was applied for all transformer models (BERT, RoBERTa, M-BERT, XLM-RoBERTa), selected based on the distribution of ACLED notes field lengths. To determine the most effective training configuration, BERT was evaluated using three optimizers under identical conditions: ADAM was run for 10 epochs with a learning rate of 2×10^{-5} and early stopping at patience 3; LAMB and ADAFACTOR were evaluated under the same epoch budget and learning rate to ensure a controlled comparison. Hyperparameter selection plays a crucial role in model convergence and classification accuracy [38], with learning rate and optimizer choice proving particularly critical in our experiments.

For multilingual evaluation, bert-base-multilingual-cased was used for M-BERT and xlm-roberta-base for XLM-RoBERTa. BERT was evaluated using three optimizers (ADAM, LAMB, and ADAFACTOR) across different temporal splits, whereas RoBERTa was fine-tuned with the AdamW optimizer for 10,000 update steps. M-BERT, pre-trained on more than 100 languages [39], and XLM-RoBERTa were employed to examine multilingual robustness and performance in low-resource language settings [37]. All models were trained on approximately 50,000 ACLED event records from 2010 to 2019, with 8000 records from 2020 reserved for evaluation under real-world class imbalance conditions. All experiments were conducted on Google Colab using a TPU v2 accelerator.

The LSTM and LLaMA inference were run on this TPU environment, while transformer-based models (BERT, RoBERTa, XLM-RoBERTa, M-BERT) were fine-tuned using the same Colab TPU v2 setup.

Table 3. (**a**) Comparison of the performance of fine-tuned traditional models, deep learning models, and zero-shot LLMs on the task of classifying political violence. (**b**) Per-class recall (%) and balanced accuracy for six representative models spanning all evaluated model families (traditional ML, zero-shot LLMs, and multilingual transformer).

(a)

No.	Models	Accuracy	Precision	Recall	F1.
	English Dataset (EN)				
1.	Decision Tree	**98%**	**0.98**	**0.76**	0.75
2.	Naïve Bayes	97%	0.73	0.54	0.56
3.	Random Forest	97%	0.8	0.62	0.64
4.	Logistic Regression	**98%**	0.90	0.73	**0.77**
5.	K-NN	96%	0.84	0.54	0.56
6.	LSTM	93%	0.19	0.30	0.20
7.	RoBERTa + ADAM	97%	0.61	0.58	0.57
8.	BERT+ ADAM	92%	0.15	0.17	0.16
9.	BERT+ ADAFACTOR	5%	0.17	0.09	0.01
10.	BERT+ LAMB	4%	0.14	0.00	0.01
11	DeepSeek (Zero-shot)	87%	0.54	0.18	0.20
12	Llama 3.2 3B instruct (Zero-shot)	55%	0.61	0.59	0.49
	Urdu Dataset (UR)				
13	M-BERT + ADAM + UR	2%	0.21	0.01	0.01
14	XLM-RoBERTa	**98%**	**0.89**	**0.84**	**0.86**

(b)

Model	Battles	Explosions	Other	Protests	Riots	Violence	Balanced Accuracy
Logistic Regression	**90.7**	91.1	46.2	**99.5**	57.6	61.6	74.4
Decision Tree	**96.0**	88.9	34.8	**99.9**	45.7	**78.8**	74.0
BERT + ADAM	0.0	0.0	0.0	**100.0**	0.0	0.0	16.7
DeepSeek R1	2.8	**94.4**	0.0	93.2	0.0	1.1	31.9
LLaMA 3.2 3B Instruct	73.0	65.9	31.6	53.5	**93.1**	46.9	60.7
XLM-RoBERTa (Urdu)	**95.2**	**96.7**	**65.4**	**99.8**	55.4	**85.6**	**83.0**

Note: Bold values indicate the best performance per metric.

The software environment consisted of Python, TensorFlow 2.9 /Keras 2.9 for the LSTM model, Hugging Face Transformers for all transformer-based and LLaMA models, scikit-learn for traditional ML models, and NLTK for text preprocessing (stopword removal, stemming, and lemmatization). The LSTM maximum sequence length was set to 512 tokens based on the 95th percentile of token counts in the dataset, and the embedding dimension was set to 64. RoBERTa was fine-tuned for 10,000 update steps, determined via validation loss monitoring with early stopping applied at a patience of 3 epochs. A batch size of 32 and learning rate of 2×10^{-5} were set following standard transformer fine-tuning guidelines [35].

To conduct zero-shot evaluation, two large language models, DeepSeek [26] and LLaMA 3.2 3B Instruct [27], were selected. DeepSeek represents a general-purpose foun-

dational model, whereas LLaMA 3.2 3B Instruct is explicitly optimized to follow classification instructions. Both models were evaluated on the same 2020 test dataset used for the fine-tuned models, employing standardized prompts. This setup ensures a fair comparison by isolating the effect of task-specific fine-tuning, as no additional training or domain adaptation was applied to the zero-shot models.

5. Performance on the Political Violence Classification Task

Table 3a summarizes model performance across all evaluation dimensions. Logistic Regression and Decision Tree both reached 98% overall accuracy; however, as the majority-class baseline (~74% accuracy, balanced accuracy = 16.7%) established in Section 4 confirms, this figure alone is not a reliable performance indicator on this dataset.

Logistic Regression achieved the best balance across all evaluation dimensions (weighted F1-score = 0.77, balanced accuracy = 74.4%), but minority-class recall remains a critical concern. For high-consequence categories such as violence against civilians (recall = 61.6%) and battles (recall = 90.7%), misclassification carries greater real-world consequence than errors in the dominant protest class, and these figures must be weighed accordingly when assessing deployment readiness.

Random Forest (97% accuracy, weighted F1 = 0.64) balanced accuracy as shown in Table 3b, and Naïve Bayes (97% accuracy, weighted F1 = 0.56) also showed stable performance, whereas K-NN produced comparatively weaker results (96% accuracy, weighted F1 = 0.56). As Table 3b illustrates, even models with near-identical overall accuracy diverge substantially on balanced accuracy and per-class recall. BERT + ADAM achieves 92% overall accuracy but a balanced accuracy of only 16.7%, equivalent to the majority-class baseline. The confusion matrices in Figure 3 reveal that while Logistic Regression and Decision Tree maintain high recall for the dominant protest class, recall for minority categories, particularly "Other" (46.2% and 34.8% respectively) and Riots, is substantially lower.

Deep learning models exhibited mixed performance. The LSTM model achieved 93% accuracy but suffered from low precision (0.19) and recall (0.30), highlighting its sensitivity to class imbalance. RoBERTa + ADAM performed relatively well, achieving 97% accuracy with an F1-score of 0.57. In contrast, BERT-based models showed strong dependence on optimizer choice. While BERT + ADAM achieved 92% accuracy, it demonstrated poor precision and recall (0.15 and 0.17, respectively). Performance dropped dramatically with alternative optimizers, as BERT + LAMB and BERT + ADAFACTOR failed to converge effectively, confirming ADAM as the only viable optimizer for this task.

The near-random performance of BERT + LAMB (4% accuracy, F1 = 0.01) and BERT + ADAFACTOR (5% accuracy, F1 = 0.01) can be attributed to two compounding factors. First, LAMB is specifically designed for large-batch distributed training scenarios and relies on gradient scaling behavior that assumes homogeneous, large-scale corpora conditions not met by the sparse, imbalanced ACLED dataset.

Second, ADAFACTOR's aggressive memory-reduction strategy eliminates the full second-moment estimate, which destabilizes convergence on small, highly skewed class distributions. Both models collapsed to predicting only the dominant class (Protests, ~74% of events), yielding near-zero recall and F1-scores across minority categories. These results reflect optimizer–dataset mismatch, consistent with the recommendation to use ADAM-family optimizers for structured, imbalanced NLP classification tasks. These discrepancies illustrate the sensitivity of deep learning models to class imbalance and optimization settings, despite their theoretical advantages [40].

All three BERT optimizer variants were trained under strictly identical conditions: same English dataset, same temporal split, same sequence length of 128 tokens, same batch size of 32, and same learning rate of 2×10^{-5}. The optimizer was the sole variable across

these three experiments. BERT + ADAM achieved 92% accuracy under these conditions while BERT + LAMB and BERT + ADAFACTOR both collapsed to near-zero performance. Both failed models assigned the overwhelming majority of predictions to the dominant Protests class, which constitutes approximately 74% of the dataset. This systematic class collapse is consistent with documented optimizer failure patterns on imbalanced datasets, distinguishable from random error by the concentrated distribution of predictions toward the majority class.

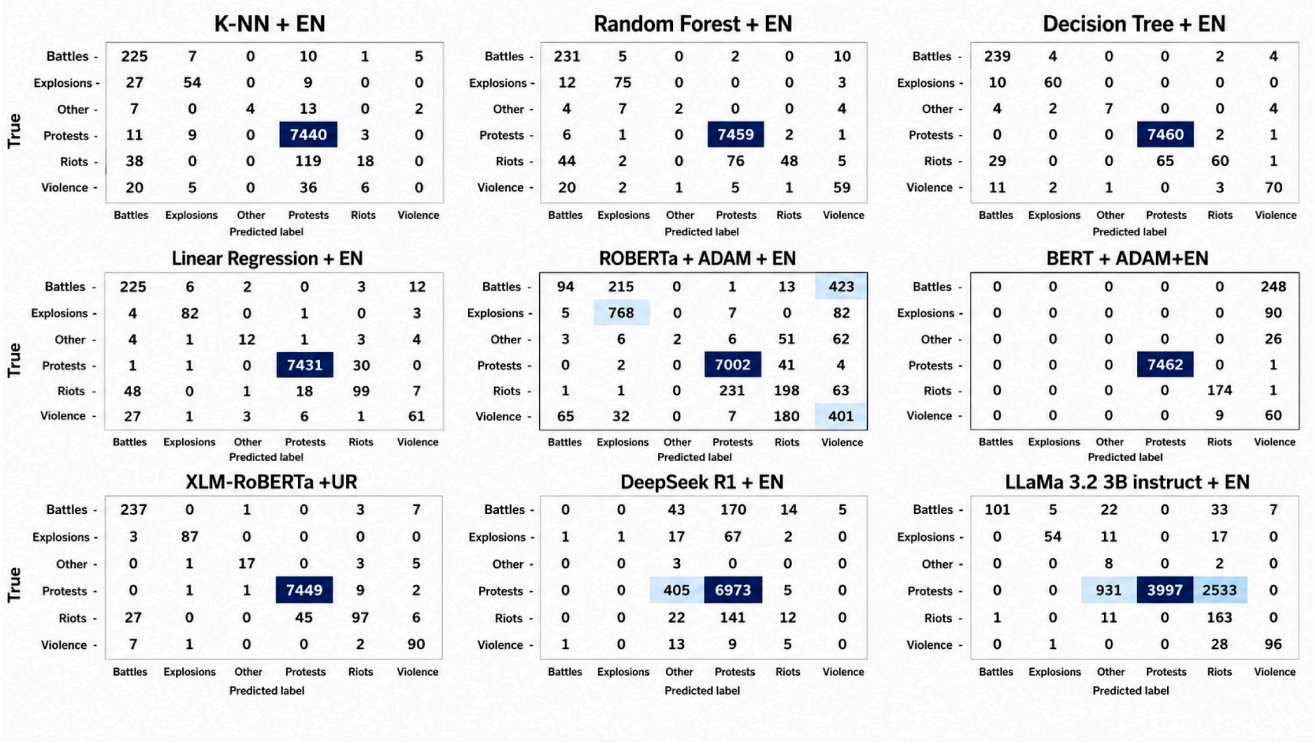


Figure 3. Confusion matrices for traditional models and deep learning models, based on all events classification performance. Color intensity represents prediction frequency: darker cells indicate higher prediction counts, while lighter cells indicate lower counts. Diagonal cells (top-left to bottom-right) represent correct classifications (True Positives) and should ideally appear darkest; off-diagonal cells represent misclassifications and should ideally appear lightest.

Zero-shot large language models performed substantially worse than fine-tuned approaches. DeepSeek achieved 87% accuracy but a low F1-score of 0.20, while LLaMA 3.2 3B Instruct reached only 55% accuracy with an F1-score of 0.49. The performance gap of 11–43% compared to fine-tuned Logistic Regression highlights that general-purpose pretraining alone is insufficient for specialized structured event classification tasks.

Multilingual models produced contrasting outcomes. M-BERT failed to perform effectively on Urdu (2% accuracy) dataset. The 2% accuracy of M-BERT on Urdu reflects a systematic failure rather than a marginal underperformance. M-BERT's Urdu vocabulary coverage is limited due to its reliance on Wikipedia-sourced multilingual pretraining, where Urdu represents less than 0.1% of the corpus. Combined with the morphologically rich and script-specific properties of Urdu text, this leads to excessive unknown-token mapping during tokenization, effectively degrading inputs to near-random feature vectors. The model consequently collapses to a degenerate prediction pattern, serving as a critical negative baseline that empirically validates the necessity of large-scale multilingual pretraining.

In contrast, XLM-R achieved 98% accuracy on Urdu, with strong precision (0.89), recall (0.84), and F1-score (0.86), demonstrating that appropriate architecture selection is critical for effective multilingual modeling.

To rule out implementation artifacts, we note that M-BERT and XLM-RoBERTa were trained using identical preprocessing pipelines, data splits, tokenization workflows, and training configurations. The only difference was the model architecture and pretraining corpus. XLM-RoBERTa achieved 98% accuracy on the same Urdu dataset under these identical conditions, providing strong controlled evidence that M-BERT's 2% accuracy reflects a genuine model–data mismatch rather than a preprocessing or implementation error. Furthermore, M-BERT's prediction pattern, collapsing entirely to the dominant class, is consistent with known failure behavior when tokenization produces predominantly unknown tokens, effectively reducing all inputs to near-identical feature vectors regardless of their actual content.

To further substantiate this explanation with empirical evidence, tokenization diagnostic analysis was conducted on 500 stratified samples from the Urdu test set using both tokenizers under identical input conditions.

Results confirmed that M-BERT produces a subword fragmentation rate of 40.58% on Urdu text, compared to 27.10% for XLM-RoBERTa. This substantially higher fragmentation rate reflects M-BERT's limited Urdu vocabulary coverage, stemming from its reliance on Wikipedia-based multilingual pretraining where Urdu constitutes less than 0.1% of the corpus. Excessive subword fragmentation degrades the representational quality of encoded inputs, producing near-identical feature vectors irrespective of actual content and explaining the observed prediction collapse. Confusion matrix analysis, presented in Figure 4, provides direct visual confirmation: M-BERT assigns the overwhelming majority of predictions across all six event categories to a single class (Battles), whereas XLM-RoBERTa demonstrates strong diagonal dominance with well-distributed classification across all six categories. Label mapping was verified to be identical across both models, with the same six-class integer encoding applied consistently. Collectively, these diagnostic checks confirm that the extreme performance divergence between M-BERT and XLM-RoBERTa reflects a genuine model–data mismatch attributable to vocabulary coverage and pretraining corpus limitations rather than any implementation or label-mapping artifact.

Overall, under the present TF-IDF-based experimental setting using short structured ACLED event descriptions, traditional machine learning models achieved stronger performance than transformer-based and zero-shot approaches, with Logistic Regression producing the most consistent results. These findings suggest that increased model complexity does not necessarily translate into improved performance for sparse structured event classification tasks represented through lexical features.

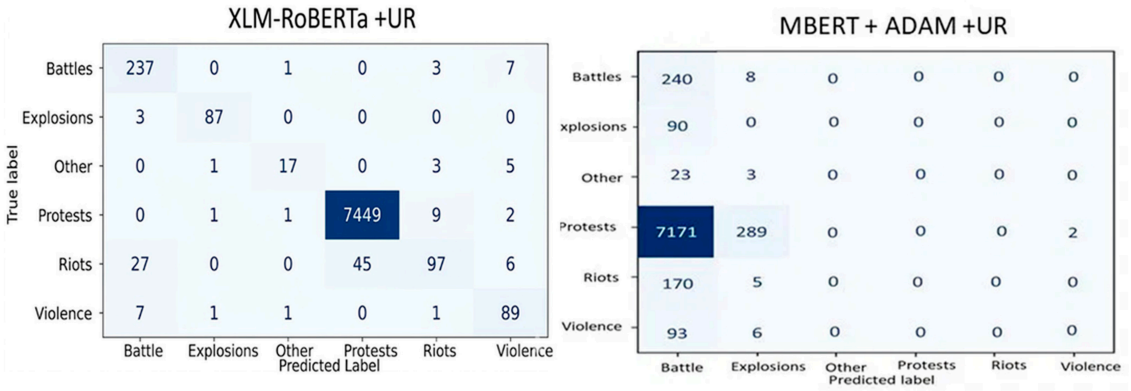


Figure 4. Confusion matrices for multilingual transformer models (M-BERT and XLM-RoBERTa) on the Urdu dataset. Color intensity represents prediction frequency: darker cells indicate higher prediction counts, while lighter cells indicate lower counts.

Comparison with Existing Studies

Compared to prior ACLED-based studies, this work extends evaluation in two respects, as shown in Table 4. While previous studies focused exclusively on English [8,13,25,41], the present study includes multilingual evaluation on Urdu. Moreover, this study compares multilingual performance (English + Urdu) and zero-shot LLMs and finds that there is an 11% performance difference (DeepSeek 87% vs. fine-tuned Logistic Regression 98%), indicating the importance of domain-specific fine-tuning. Moreover, our Twitter based analysis of the Kashmir Article 370 crisis (57,700 tweets) provides exploratory discourse indicators during a major crisis event, a dimension not addressed by methods relying solely on historical data. Because the best Twitter classification accuracy (55%) remains below the 69% majority-class baseline, this component is positioned strictly as an exploratory discourse analysis exercise and is deliberately set apart from the ACLED classification findings. It illustrates the framework's extensibility to unstructured social media text but does not contribute to the primary classification evaluation.

Table 4. Performance comparison with ACLED-based political violence classification studies.

Study	Dataset	Models	Accuracy	Language
[8]	ACLED + Newspaper (1997–2014)	Lasso Regression, NLP	80%	EN
[25]	ACLED + UCDP + PRIO	RF Ensemble	0.85% (state) and 0.82% (non-state)	EN
[13]	ACLED + Satellite (Landsat, Sentinel)	CNN (DL)	97%	EN
Our approach	**ACLED (2010–2020) + Twitter**	**ML + DL + LLMs**	**98% (balanced acc. 74.4%; macro-F1: see Table 3b)**	**EN + UR**

Note: Bold values indicate the best performance.

6. Exploratory Case Study: Analyzing Public Discourse During the Article 370 Crisis Using Twitter Data

As a negative result extension, the framework was applied to 57,700 English tweets collected during the Article 370 crisis (August–September 2019) to test applicability beyond structured ACLED data. Tweets from verified political, media, and intergovernmental accounts were annotated into four stance categories (In Favor of Kashmir, Against India, In Favor of India, Against Kashmir; $\kappa = 0.78$). The best-performing model (Logistic Regression) achieved 55% accuracy, which falls below the 69% majority-class baseline confirming that the framework does not transfer effectively to unstructured social media text under these conditions [42,43].

This negative result is included not as a validated classification contribution but to motivate future work: effective social media stance classification would require representative sampling, balanced training data, and domain-adapted models. The primary findings of this study rest entirely on the ACLED structured-event results reported in Section 5.

Data collection used keyword-based filtering ("Article370", "Kashmir", "370Revoked") targeting verified political, media, and intergovernmental accounts (n = 250) between 5 August and 15 September 2019. From 62,543 raw tweets, 4843 duplicates were removed via tweet ID matching, and 1523 tweets from potential bot accounts (>50 tweets/day) were excluded, yielding 57,700 tweets. Retweets were retained as they reflect discourse amplification. All tweets were in English, verified via Twitter API language metadata. Data collection complied with Twitter's Terms of Service using publicly accessible accounts.

The tweets were manually annotated into four categories: In Favor of Kashmir, In Favor of India, Against India, and Against Kashmir. To ensure objectivity and minimize annotation bias, two independent annotators were recruited with complementary expertise: one with a background in Computer Science specializing in Natural Language Pro-

cessing, and the other with expertise in Communication and Media Studies focusing on political discourse analysis. This interdisciplinary pairing enabled annotations to be evaluated from both technical and socio-political perspectives, thereby reducing the risk of domain-specific bias.

The annotated dataset shows pronounced imbalance: 69% "In Favor of Kashmir" (39,803), 19% "Against India" (10,963), 8% "In Favor of India" (4616), and 4% "Against Kashmir" (2318), reflecting organic polarization of elite political discourse during the crisis. Prior to the annotation task, detailed labeling guidelines were developed, defining each category with concrete examples to standardize interpretation.

Each tweet was independently labeled by both annotators, and inter-annotator agreement was assessed using Cohen's Kappa coefficient, which yielded a score of $\kappa = 0.78$, indicating substantial agreement.

In cases of disagreement, a structured discussion protocol was employed, wherein annotators reviewed conflicting labels and reached consensus through dialog. This approach ensured that final annotations reflected balanced judgment rather than individual bias, thereby enhancing the reliability and validity of the labeled dataset.

Furthermore, the word frequency analysis was done to detect dominant terms in the tweets data and extract meaningful insights. The 10 most frequent words in the dataset are highlighted in Table 5. In this analysis, words like 'Pakistan', 'India', 'Minister', and 'Kashmir' reflect the geopolitical focus of the data. TF-IDF was used to represent high-frequency terms such as 'Pakistan' and 'Kashmir' as feature vectors for the analysis of semantic and contextual sentiment. Through this process, critical terms and their definitions were obtained, which uncovered the emotional range of our dataset.

Table 5. Word frequency in the Twitter dataset.

Word	No. of Occurrences
Pakistan	3165
India	2829
Minister	1963
Kashmir	1583
Indian	1175
Today	1062
Khan	1037
Prime	992
People	902
President	774

This case study explores discourse patterns in social media content during a politically sensitive event, with results interpreted strictly as illustrative indicators rather than validated classifications. Moreover, experiments were conducted with machine learning models including Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF) and K-Nearest Neighbors (K-NN) to identify the most appropriate model to classify a tweet into four categories such as 'In favor of Kashmir,' 'In favor of India,' 'Against India,' and 'Against Kashmir'. The results, as shown in Table 6, were that Logistic Regression achieved the best performance (55% accuracy) and was selected for final predictions. However, these modest results reflect the inherent difficulty of four-way stance classification on politically polarized content.

As shown in Table 6, several models exhibit severely low recall for minority classes: "In Favor of India" achieves only 0.05 recall with Logistic Regression, indicating that 95% of this minority class is misclassified, primarily into the dominant "In Favor of Kashmir" category. Similarly, "Against Kashmir" shows recall of 0.32, reflecting the challenge of identifying underrepresented stances. The severe class imbalance (69% majority class) means that a naïve baseline predicting only "In Favor of Kashmir" would achieve 69% accuracy, contextualizing the 55% result as below-baseline performance that necessitates caution in interpretation. Given the below-baseline performance, these outputs should be treated as coarse-grained discourse indicators only and not as reliable stance classifications.

Table 6. Multi-class sentiment classification for Article 370 tweets using traditional machine learning models.

Models	In favor of Kashmir			In favor of India			Against India			Against Kashmir			Acc
	Pre.	Rec.	F1.	Pre.	Rec.	F1.	Pre.	Rec.	F1.	Pre.	Rec.	F1.	
NB	0.61	0.79	0.69	0.00	0.00	0.00	0.58	0.67	0.62	0.45	0.25	0.32	48%
LR	0.72	0.80	0.75	1.00	0.05	0.10	0.60	0.70	0.64	0.46	0.32	0.37	55%
SVM	0.76	0.76	0.76	0.50	0.17	0.26	0.62	0.70	0.66	0.4	0.38	0.41	53%
RF	0.69	0.68	0.69	0.30	0.07	0.12	0.57	0.57	0.57	0.38	0.34	0.36	44%
K-NN	0.29	0.73	0.42	0.12	0.12	0.12	0.25	0.10	0.14	0.24	0.28	0.26	25%

Analysis revealed that discourse was dominated by pro-Kashmir sentiment (69%), with "Against India" (19%), "In Favor of India" (8%), and "Against Kashmir" (4%) constituting minority positions.

These findings, represented in Figure 5, reflect the thematic orientation of public debate among sampled political and media accounts during the crisis period.

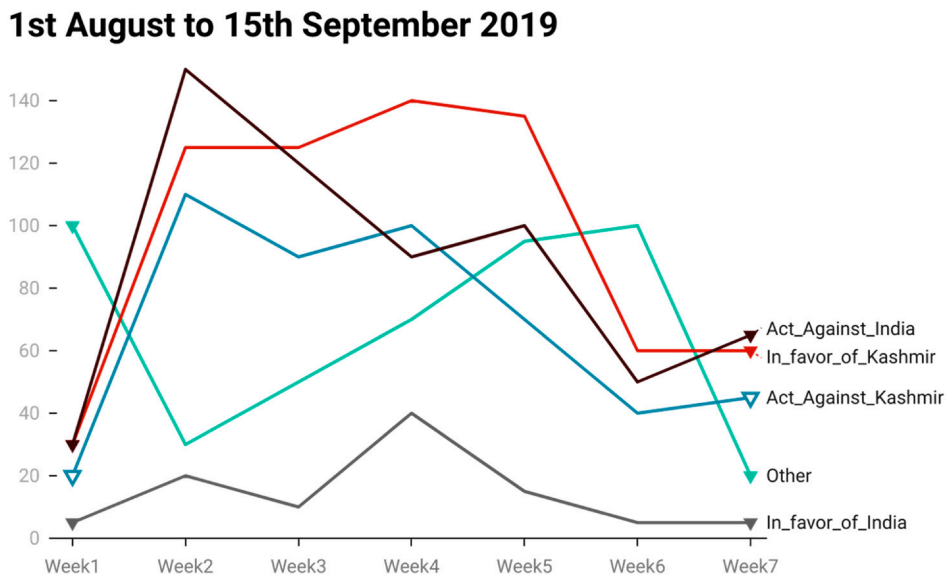


Figure 5. Weekly sentiment trends indicating public support for the restoration of Kashmir with a possibility of the restoration of Article 370.

However, several methodological limitations constrain interpretation: (1) the sample is not representative of general public opinion, being drawn exclusively from political elites and media rather than ordinary citizens; (2) severe class imbalance and modest classification performance (55% accuracy, below the 69% naïve baseline) limit reliable tracking of minority stances; (3) the temporal snapshot (August–September 2019) captures crisis-period discourse but cannot establish causal relationships with subsequent policy

developments; and (4) platform-specific biases (Twitter's user demographics, algorithmic amplification, coordinated campaigns) may distort apparent sentiment distributions. This case study illustrates the challenges of applying classification frameworks to unstructured social media text. The below-baseline performance confirms that the Twitter component does not constitute a validated classification framework and should be interpreted as an exploratory extension only, included to illustrate applicability beyond structured event data rather than to claim analytical parity with the ACLED results. Future work would require representative sampling, balanced training data, and continuous model recalibration to approach operational deployment for policy monitoring.

7. Discussion

The superior performance of traditional machine learning models can be primarily attributed to their strong alignment with the structural characteristics of the ACLED dataset and the assumptions underlying linear classification. ACLED is an event-coded dataset comprising discrete categorical labels, sparse and concise textual descriptions, and well-defined structured metadata, including event type, sub-event, geographic location, and fatality counts. These properties are well suited to classifiers such as Logistic Regression and Decision Trees, which are specifically designed to operate effectively in high-dimensional, sparse feature spaces generated through TF–IDF representations. In contrast to free-form narrative corpora, ACLED does not necessitate deep semantic abstraction; instead, accurate classification depends on the identification of consistent lexical cues and categorical regularities. As a result, increasing model complexity does not inherently translate into improved predictive performance in this setting. A further caveat is that these results are conditioned on TF-IDF features from short ACLED notes. Had models been applied to longer conflict narratives such as full news articles, transformer-based architectures might perform substantially differently.

The contrasting performance between ACLED (98% accuracy, 74.4% balanced accuracy) and the Twitter extension (55%, below the 69% baseline) confirms that effective classification depends fundamentally on data structure, annotation quality, and class balance rather than model choice alone. This contrast motivates future work on domain adaptation for unstructured social media text.

As reported in Section 5, overall accuracy figures near 98% are substantially inflated by majority-class dominance. Per-class recall for high-consequence categories such as violence against civilians and battles warrants particular attention, as detailed in Table 3b. Logistic Regression achieved the strongest balanced accuracy among traditional models (74.4%), yet minority-class recall for 'Other' (46.2%) and Riots (57.6%) remains limited, as detailed in Table 3b. Confusion matrix analysis indicates that traditional machine learning models preserve more consistent recall across event categories, whereas deep learning architectures show greater tendency toward majority-class prediction under imbalanced conditions.

Transformer-based architectures were found to be highly sensitive to both optimization strategies and dataset scale. While BERT fine-tuned with the ADAM optimizer achieved moderate and stable performance, alternative optimizers such as LAMB and ADAFACTOR failed to converge, producing near-random classification outcomes. This instability can be attributed to the mismatch between large-parameter models and relatively small, highly imbalanced datasets, where aggressive adaptive optimization can exacerbate gradient instability. Moreover, the textual content in ACLED is typically short, factual, and repetitive, which limits the advantages offered by contextual representation learning. These findings suggest that, for structured event classification tasks, deep learning models may experience over-parameterization, thereby diminishing their practical effectiveness despite their theoretical representational capacity. The controlled experimental

design, with the optimizer as the sole variable, provides ablation-level evidence that the observed failures are attributable to optimizer–dataset mismatch.

The divergent performance of multilingual models highlights the significance of pretraining data diversity and subword modeling in low-resource language settings. Multilingual BERT, which is predominantly trained on Wikipedia, offers limited lexical coverage and morphological representation for Urdu. In contrast, XLM-RoBERTa is pretrained on a large-scale CommonCrawl corpus, enabling substantially richer subword representations and improved handling of non-Latin scripts. This difference in pretraining strategy and data scale is consistent with the substantial performance gap observed between the two models, suggesting that multilingual capability should not be inferred solely from model nomenclature.

The multilingual component of this study employs a machine-translated version of the English ACLED corpus as the Urdu evaluation dataset. This design provides a controlled indication of cross-lingual transfer under translation-based conditions. The results suggest that XLM-RoBERTa is better suited to this setting than M-BERT, though generalization to naturally occurring Urdu text remains to be established. The results should be interpreted within this experimental scope, and evaluation on independently collected naturally occurring Urdu political-event text represents a valuable extension recommended for future research.

In this experimental setting, zero-shot LLMs underperformed fine-tuned approaches, particularly in recall and F1-score. Although these models possess strong general-purpose language understanding capabilities, political violence event classification constitutes a schema-based classification task rather than an open-ended semantic reasoning problem [44]. The event labels defined in ACLED correspond to ontological categories with precise and domain-specific definitions, which are not reliably inferred through prompt-based natural language instructions.

Furthermore, zero-shot inference lacks both calibrated probability estimates and domain-specific adaptation, resulting in inconsistent performance on minority event categories. These findings suggest that, within this experimental setting, zero-shot LLMs underperform fine-tuned approaches on structured event classification tasks, though their suitability across other datasets and task formulations remains an open empirical question.

A further limitation concerns geographic generalizability. The temporal split employed here guards against data leakage across time but does not validate cross-regional performance. Because ACLED event descriptions often contain country-specific lexical patterns, actor names, place names, and conflict terminology specific to particular regions, a model trained on data from one set of countries may not generalize equally well to new geographic contexts. Leave-one-country-out or leave-one-region-out evaluation would provide a stronger test of external validity. We recommend this as a priority for future work, particularly when deploying classification frameworks for conflict classification in previously unseen regions.

8. Conclusions

This study systematically evaluated 12 machine learning models for political violence event classification using the ACLED dataset (2010–2020), complemented by a Twitter-based case study. Traditional machine learning models, particularly Logistic Regression and Decision Trees, demonstrated superior performance as measured by macro-F1 and balanced accuracy in the primary evaluation metrics given the dataset's pronounced class imbalance. While overall accuracy reached 98%, this figure must be interpreted against a majority-class baseline of ~74% overall accuracy (balanced accuracy = 16.7%, macro-F1 $\approx$ 0.13). Logistic Regression achieved a balanced accuracy of 74.4% and a weighted F1-score of 0.77, representing the strongest per-class performance among all

models evaluated, yet minority-class recall for categories such as "Other" (46.2%) and Riots (57.6%) remains limited, as shown in Table 3b. Deep learning architectures exhibited optimizer sensitivity and collapsed to majority-class prediction under imbalanced conditions (BERT + ADAM: balanced accuracy = 16.7%), while zero-shot large language models underperformed fine-tuned approaches, suggesting that general-purpose pre-training may not substitute domain-specific optimization on structured event classification tasks of this type. Multilingual evaluation revealed architecture-dependent disparities: XLM-RoBERTa achieved 98% accuracy and 83.0% balanced accuracy on machine-translated Urdu, whereas Multilingual BERT failed with a balanced accuracy of 16.7% (equivalent to the majority-class baseline), highlighting the importance of pre-training data diversity and subword modeling for low-resource languages.

The Twitter case study was included as an exploratory extension to illustrate framework applicability to unstructured social media text; given the below-baseline performance, no validated classification claims are drawn from this component. Importantly, with 55% accuracy falling below the 69% majority class baseline, this component does not constitute a predictive or deployment-ready monitoring tool and should be interpreted as a proof of concept only. For analytical applications, traditional models applied to structured ACLED data offer classification reliability, transparent decision-making, and CPU-based inference without specialized infrastructure, making them suitable for resource-constrained research and monitoring contexts.

A key limitation is that ACLED notes are short structured summaries, which structurally favors TF-IDF-based lexical models. A further limitation is that geographic out-of-sample validation was not performed; results are reported for a single temporally split evaluation and have not been validated across held-out countries or regions. Readers should therefore not interpret the reported performance as evidence of cross-regional generalizability; leave-one-country-out or leave-one-region-out experiments are needed before deploying this framework in novel geographic contexts. This finding should not be generalized to scenarios involving richer or longer textual sources. Future research should explore hybrid architectures combining traditional ML with transformers for textual content, address class imbalance through focal loss or SMOTE, integrate real-time data streams, and validate geographic generalization through leave-one-country-out or leave-one-region-out experiments to assess whether classification performance holds across unseen geographic contexts. This study provides a methodological benchmark indicating that, on structured coded event data of this type, aligning model capabilities with dataset characteristics may matter more than architectural complexity.

Author Contributions: Conceptualization, U.B., S.A.R., F.M., and M.G.A.M.; methodology, U.B., F.M., and M.I. (Muhammad Imran).; software, U.B. and F.M.; validation, U.B., F.M., S.A.R., and M.I. (Muhammad Imran).; formal analysis, U.B., M.I., M.G.A.M., and M.I. (Muddessar Iqbal).; investigation, U.B., F.M., and S.I.N.; resources, S.A.R., M.I. (Muddessar Iqbal), and M.G.A.M.; data curation, U.B., F.M., and M.I. (Muhammad Imran).; writing—original draft preparation, U.B. and F.M.; writing—review and editing, S.A.R., S.I.N., M.G.A.M., M.I. (Muddesar Iqbal)., and M.G.A.M.; visualization, U.B., and F.M.; supervision, S.A.R.; project administration, S.A.R., and M.I. (Muhammad Imran). All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Prince Sultan University.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research was conducted using publicly available secondary datasets, including the ACLED dataset and anonymized Twitter data collected through the Twitter API. This study did not involve direct interaction with human participants, collection of private identifiable information, clinical data, or any intervention requiring institutional ethical oversight.

Informed Consent Statement: Not applicable. This study did not involve direct participation of human subjects, as the analysis was conducted using publicly available secondary datasets and anonymized social media data.

Data Availability Statement: The data used in this study are publicly available. The Armed Conflict Location and Event Data Project (ACLED) dataset can be accessed from the official ACLED website https://acleddata.com/data/ (accessed on 1 January 2022). The Twitter dataset used in the case study was collected via the Twitter API and is subject to platform data-sharing policies. Processed data and code used for analysis are available from the corresponding author upon reasonable request.

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

References

1. Halder, R.; Chatterjee, R. Cnn-bilstm model for violence detection in smart surveillance. *SN Comput. Sci.* **2020**, *1*, 201. [CrossRef]
2. Jenkins, J.C.; Taylor, C.L.; Abbott, M.; Maher, T.V.; Peterson, L. *Global Conflict Data: Introducing the World Handbook of Political Indicators IV*; The Ohio State University: Columbus, OH, USA, 2018.
3. Zahid, M.S.; Jamil, U. Data-driven machine learning techniques for fuel economy prediction in sustainable transportation systems. *Green Energy Intell. Transp.* **2026**, *5*, 100303. [CrossRef]
4. Ezen-Can, A. A comparison of lstm and bert for small corpus. *arXiv* **2020**, arXiv:2009.05451.
5. Pettersson, T.; Eck, K. Organized violence, 1989–2017. *J. Peace Res.* **2018**, *55*, 535–547. [CrossRef]
6. Anastasopoulos, L.J.; Williams, J.R. A scalable machine learning approach for measuring violent and peaceful forms of political protest participation with social media data. *PLoS ONE* **2019**, *14*, e0212834. [CrossRef] [PubMed]
7. Chadefaux, T. Conflict forecasting and its limits. *Data Sci.* **2017**, *1*, 7–17. [CrossRef]
8. Mueller, H.; Rauh, C. The hard problem of prediction for conflict prevention. *J. Eur. Econ. Assoc.* **2022**, *20*, 2440–2467. [CrossRef]
9. Xu, Z.; Lin, Y.; Cai, H.; Zhang, W.; Shi, J.; Situ, L. Risk assessment and categorization of terrorist attacks based on the global terrorism database from 1970 to 2020. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1103. [CrossRef]
10. Badshah, A.; Daud, A.; Alharbey, R.; Banjar, A.; Bukhari, A.; Alshemaimri, B. Big data applications: Overview, challenges and future. *Artif. Intell. Rev.* **2024**, *57*, 290. [CrossRef]
11. Deng, S.; Ning, Y. A survey on societal event forecasting with deep learning. *arXiv* **2021**, arXiv:2112.06345.
12. de Mesquita, E.B.; Fair, C.C.; Jordan, J.; Rais, R.B.; Shapiro, J.N. Measuring political violence in pakistan: Insights from the bfrs dataset. *Confl. Manag. Peace Sci.* **2015**, *32*, 536–558.
13. Warnke, S.; Runfola, D. Predicting protests and riots in urban environments with satellite imagery and deep learning. *Trans. GIS* **2024**, *28*, 2309–2327. [CrossRef]
14. Murphy, M.; Sharpe, E.; Huang, K. The promise of machine learning in violent conflict forecasting. *Data Policy* **2024**, *6*, e35. [CrossRef]
15. Chen, Z.; Ma, W.; Li, Y.; Guo, W.; Wang, S.; Zhang, W.; Chen, Y. Using machine learning to estimate the incidence rate of intimate partner violence. *Sci. Rep.* **2023**, *13*, 5533. [CrossRef] [PubMed]
16. Chiba, D.; Gleditsch, K.S. The shape of things to come? expanding the inequality and grievance model for civil war forecasts with event data. *J. Peace Res.* **2017**, *54*, 275–297. [CrossRef]
17. Saeed, Z.; Abbasi, R.A.; Maqbool, O.; Sadaf, A.; Razzak, I.; Daud, A.; Aljohani, N.R.; Xu, G. What's happening around the world? a survey and framework on event detection techniques on Twitter. *J. Grid Comput.* **2019**, *17*, 279–312. [CrossRef]
18. Bahrami, M.; Findik, Y.; Bozkaya, B.; Balcisoy, S. Twitter reveals: Using Twitter analytics to predict public protests. *arXiv* **2018**, arXiv:1805.00358.
19. Hegre, H.; Allansson, M.; Basedau, M.; Colaresi, M.P.; Croicu, M.; Fjelde, H.; Hoyles, F.; Hultman, L.; Högbladh, S.; Jansen, R.; et al. ViEWS: A political violence early-warning system. *J. Peace Res.* **2019**, *56*, 155–174. [CrossRef]
20. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 39. [CrossRef]
21. Khan, W.; Daud, A.; Khan, K.; Muhammad, S.; Haq, R. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Nat. Lang. Process. J.* **2023**, *4*, 100026. [CrossRef]
22. Naz, A.; Khan, H.U.; Alesawi, S.; Abouola, O.I.; Daud, A.; Ramzan, M. AI knows you: Deep learning model for prediction of extroversion personality trait. *IEEE Access* **2024**, *12*, 159152–159175. [CrossRef]
23. Hayat, M.K.; Daud, A.; Alshdadi, A.A.; Banjar, A.; Abbasi, R.A.; Bao, Y.; Dawood, H. Towards deep learning prospects: Insights for social media analytics. *IEEE Access* **2019**, *7*, 36958–36979. [CrossRef]
24. Lynam, T.; Zapata, M.; Hegre, H.; Bell, C.; Besaw, C. Early warning and predictive analytic systems in conflict contexts: Insights from the field. *Civ. Wars* **2023**, *26*, 401–429. [CrossRef]

25. Raleigh, C.; Dowd, C. *Armed Conflict Location and Event Data Project (ACLED) Codebook*; ACLED: Madison, WI, USA, 2015.
26. Deng, Z.; Ma, W.; Han, Q.-L.; Zhou, W.; Zhu, X.; Wen, S.; Xiang, Y. Exploring deepseek: A survey on advances, applications, challenges and future directions. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 872–893. [CrossRef]
27. Tan, Y.; Wu, B.; Cao, J.; Jiang, B. Llama-utp: Knowledge-guided expert mixture for analyzing uncertain tax positions. *IEEE Access* **2025**, *13*, 90637–90650. [CrossRef]
28. Kratzert, F.; Gauch, M.; Klotz, D.; Nearing, G. Hess opinions: Never train a long short-term memory (lstm) network on a single basin. *Hydrol. Earth Syst. Sci.* **2024**, *28*, 4187–4201. [CrossRef]
29. Huang, S.; Cai, N.; Pacheco, P.P.; Narrandes, S.; Wang, Y.; Xu, W. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genom. Proteom.* **2018**, *15*, 41–51. [CrossRef]
30. Rigatti, S.J. Random forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef] [PubMed]
31. Xu, S. Bayesian naïve bayes classifiers to text classification. *J. Inf. Sci.* **2018**, *44*, 48–59.
32. Charbuty, B.; Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
33. Boateng, E.Y.; Abaye, D.A. A review of the logistic regression model with emphasis on medical research. *J. Data Anal. Inf. Process.* **2019**, *7*, 190. [CrossRef]
34. Niknejad, M. Deep learning in business analytics: A clash of expectations and reality. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100146.
35. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Bidirectional encoder representations from transformers. *arXiv* **2018**, arXiv:1810.04805.
36. Delobelle, P.; Winters, T.; Berendt, B. Robbert: A dutch RoBERTa-based language model. *arXiv* **2020**, arXiv:2001.06286.
37. Prytula, M. Fine-tuning bert, distilbert, xlm-roberta and ukr-roberta models for sentiment analysis of ukrainian language reviews. *Mach. Learn.* **2024**, *3*, 85–97. [CrossRef]
38. Zhan, T. Hyper-parameter tuning in deep neural network learning. In *Artificial Intelligence and Applications*; AIRCC Publishing Corporation: Chennai, India, 2022; pp. 95–101.
39. Rahman, M.A.; Shahriar, H.; Wu, F.; Cuzzocrea, A. Applying pre-trained multilingual bert in embeddings for improved malicious prompt injection attacks detection. In *Proceedings of the 2024 2nd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*; IEEE: New York, NY, USA, 2024; pp. 1–7.
40. Juba, B.; Le, H.S. Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2019; Volume 33, pp. 4039–4048.
41. Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. Deepseek-v3 technical report. *arXiv* **2024**, arXiv:2412.19437.
42. Mukherjee, S.; Chatterjee, A.; Dass, S. Deciphering the realities of deep learning in business analytics: A bibliometric analysis. In *Intelligent Optimization Techniques for Business Analytics*; IGI Global: Hershey, PA, USA, 2024; pp. 77–103.
43. Nasuto, A.; Rowe, F. Understanding anti-immigration sentiment spreading on Twitter. *PLoS ONE* **2024**, *19*, e0307917. [CrossRef] [PubMed]
44. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.