Forecasting Online Negativity Spikes with Multilingual Transformers for Strategic Decision-Making

Rowan Martnishn and **Viraj Chawda** and **Vishal Green** and **Julia Brady** and **Ashley Lauw**
Ria Rathi and Shravan Athikinasetti and Varun Kadari and Zachary Miller and Nikhil Badlani
Sentivity.ai
Blacksburg, Virginia, USA
Contact: rowan@sentivity.ai

Abstract

Social media platforms like Reddit, YouTube, and Instagram amplify rapid dissemination of negative sentiment, potentially causing harm and fostering extremist discourse. This paper addresses the NLP challenge of predicting sudden spikes in negative sentiment by fine-tuning multilingual transformer models. We present a structured pipeline emphasizing linguistic feature extraction and temporal modeling. Our experimental results, obtained from extensive Reddit, YouTube, and Instagram data, demonstrate improved forecasting accuracy over baseline methods. Ethical considerations and implications for deployment in social media moderation are thoroughly discussed. The system includes user-centric interactive features such as real-time filtering dashboards, customizable negativity thresholds, and forecasting analytics, providing actionable insights for preventative content moderation. Given its real-time deployment potential and cross-platform applicability, our system offers actionable insights for proactive content moderation.

1 Introduction

Social media platforms play a critical role in communication, but frequently propagate toxic narratives. Reddit, YouTube, and Instagram, in particular, see frequent bursts of negativity, affecting user well-being and broader social discourse. Accurately forecasting such spikes enables proactive moderation and intervention. This paper introduces an NLP-centric solution: forecasting sentiment spikes through customized transformer models tailored to these platforms' unique linguistic landscape. Our implemented pipeline combines a fine-tuned multilingual transformer model (cardiffnlp/xlm-twitter-politics-sentiment) and an XGBoost regression model enhanced by Gaussian smoothing. To analyze narrative clusters, we applied HDBSCAN clustering combined with OpenAI embeddings.

Our contributions are as follows:

- A refined adaptation of the cardiffnlp/xlm-twitter-politics-sentiment model for social media-specific language;
- A linguistically informed feature extraction pipeline that captures early signals of negativity spikes;
- A robust forecasting model combining neural networks and ensemble learning for multi-day sentiment prediction;
- A rigorous evaluation framework cross-referenced with real-world events and historical spikes;
- Comprehensive ethical and practical framework for safe deployment, addressing potential misuse, user privacy, and algorithmic bias.

2 Related Work

Sentiment analysis traditionally focused on static text classification; recent studies increasingly integrate temporal dynamics for real-time online prediction. Transformer-based models, particularly multilingual variants, show significant social media promise. Our approach advances this by combining fine-grained linguistic analysis with temporal forecasting, addressing existing literature gaps.

2.1 Toxicity and Negativity Spike Prediction

Early efforts to identify harmful online discussions relied on burst detection (Papegnies et al., 2017), sequence modeling for civil-to-toxic shifts (Zhang et al., 2018), or LSTM-based models for hate speech temporal dynamics (Mathew et al., 2021). While highlighting temporal/conversational

Proceedings of Recent Advances in Natural Language Processing,pages 704–710
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-081

modeling's importance, these focus on toxicity *detection*, rarely applying to cross-platform contexts or prioritizing sentiment-driven *early warning systems* for proactive intervention, a core focus here.

2.2 Multilingual Sentiment and Offense Detection

Transformer-based models revolutionized multilingual NLP. Models like (Sigurbergsson and Derczynski, 2020) show strong offensive language detection capabilities across languages. XLM-R's success in OffensEval (Zampieri et al., 2020) further underscores multilingual transformers' power for social media toxicity. However, studies like (Chiril et al., 2020) (FRENK) reveal persistent cross-language performance gaps, indicating a need for robust, generalizable solutions. While effective for toxicity, these works often rely on static, task-specific datasets or lack the real-time generalizability across diverse social contexts and content types our system aims for.

2.3 Content Moderation and Social Media Dynamics

Research into social media dynamics explores content moderation policy effects. Studies like (Jhaver et al., 2019) investigate how Reddit moderation shapes user engagement/community fairness. (Ribeiro et al., 2021) conducted longitudinal studies on deplatforming ripple effects, while (Faddoul et al., 2020) used network analysis to map conspiracy narrative propagation on platforms like Reddit/YouTube. These works are crucial for understanding moderation's social effects and platform-specific behaviors. However, they typically focus on analyzing existing dynamics or retrospective intervention impact, stopping short of proactively forecasting harmful content surges.

2.4 Platform-Specific Analysis: Beyond Twitter

While Twitter is a primary sentiment/toxicity focus, research also delves into other platforms. (Zhang and Davidson, 2020) explored Reddit-specific BERT models for toxicity. Studies like (Thelwall et al., 2012) analyzed YouTube comment polarity/clustering, and (Gao et al., 2018) focused on Instagram for multimodal hate speech using joint embeddings. These platform-specific analyses reveal unique linguistic/behavioral traits. However, they often lack a unified, cross-platform predictive framework, typically functioning as static classifiers rather than proactive spike forecasting systems.

2.5 Our Work's Contribution and Gaps Addressed

Our research distinctively advances the field by primarily forecasting future negative sentiment surges, rather than merely detecting them. Unlike prior snapshot classifiers, our approach predicts upcoming volatility. We introduce a socially-trained, cross-platform system, fine-tuned on social media-native discourse across Reddit, YouTube, and Instagram content. Crucially, instead of binary classification, our model produces a single continuous negativity score, inherently better suited for temporal modeling and volatility analysis to identify sudden shifts.

This work addresses several critical gaps in existing literature:

- No existing model reliably predicts cross-platform, generalizable, real-time negativity spikes.
- Current literature lacks tools proactively flagging broad sentiment trends *before* crises escalate, especially across multiple content modalities/diverse community cultures.
- Most existing multilingual/content moderation models remain platform-bound, task-specific, or retrospective, limiting proactive intervention utility.

Integrating a fine-tuned multilingual transformer with an XGBoost regression model and linguistically-informed feature extraction, our pipeline offers a novel, practical solution for proactive content moderation.

3 Dataset and Preprocessing

We used social media data from diverse forums collected over six-month periods (three months before/after) each of four major sociopolitical events: the Capitol Insurrection, the Supreme Court's reversal of *Roe v. Wade*, George Floyd's death, and COVID-19 lockdowns. While Reddit served as our core dataset for sociopolitical event evaluation, we included parallel Instagram captions and YouTube video descriptions for generalizability. All text data underwent a unified sentiment preprocessing pipeline, which involved:

3.1 Data Collection

We collected approximately 2 million posts/comments via various APIs, ensuring representation from politically charged, mental health, and general forums. For comprehensive representation, we used stratified sampling, selecting forums known for diverse discourse (Zhuravskaya et al., 2020; DiGrazia et al., 2013; Dictionary, 2025; Association, 2004). APIs provided efficient data retrieval, yielding post/comment text and metadata (timestamps, author information, forum affiliation (SubredditStats.com, 2025)).

3.2 Text Cleaning

Text standardization involved removing URLs, special characters, emojis, and mentions, plus normalizing punctuation/capitalization. Raw social media text often contains noise hindering sentiment analysis; thus, our cleaning pipeline standardized text and removed disruptive elements:

- **URLs:** Removed to eliminate external links irrelevant to sentiment.
- **Special characters:** Non-alphanumeric symbols removed to simplify text.
- **Emojis:** Removed due to potential ambiguity/inconsistency in sentiment analysis.
- **Mentions:** User mentions (e.g., u/username) removed to avoid bias.
- **Punctuation/Capitalization:** Standardized for consistent text format, reducing variations not affecting underlying sentiment.

3.3 Temporal Aggregation

Hourly aggregation created continuous sentiment time series for robust temporal modeling. To analyze trends, we aggregated post/comment sentiment scores into hourly bins, transforming discrete data into a continuous time series of average hourly sentiment. This aggregation is crucial for identifying sentiment patterns, fluctuations, and applying time series forecasting techniques. Hourly aggregation balanced capturing fine-grained dynamics with noise reduction; finer aggregation (e.g., minute-level) risks excessive noise, while coarser (e.g., daily) may obscure short-term trends.

3.4 Sentiment Labeling

We fine-tuned the transformer model on manually annotated data. Posts exceeding a determined negativity threshold were labeled ”negative”, with hourly bins flagged as spikes based on aggregated sentiment. Spikes replicated trends observed on chosen sociopolitical event days. For model training, we manually annotated a dataset, carefully assigning sentiment labels considering linguistic context/nuances. An experimentally validated negativity threshold ensured accurate sentiment reflection. Hourly bins identified periods of high negativity for spike forecasting. Manual annotation is crucial for model quality, providing labeled examples for accurate classification of unseen posts.

3.5 Data Splitting and Class Balance

Our dataset was chronologically split for evaluation to prevent data leakage, with each sociopolitical event's date serving as a natural cutoff. Data three months prior to each event was used for training/validation, and three months post-event for testing. Within these segments, we applied an 80% training, 10% validation, and 10% testing proportional random split. Given spike events are naturally less frequent, we addressed class imbalance by undersampling non-spike days for balanced representation. A classifier further filtered for ’usable’ posts, focusing on politics/relevant topics, reducing irrelevant posts far from negativity peaks and optimizing the dataset for spike detection. This strategy ensured clear training signals for rare, critical spike events.

4 Model Architecture

We adapted cardiffnlp/xlm-twitter-politics-sentiment, a multilingual transformer originally pre-trained on political discourse, to score sentiment across Reddit, YouTube, and Instagram posts after preprocessing and linguistic annotation:

4.1 Fine-Tuning Procedure

The cardiffnlp/xlm-twitter-politics-sentiment model(CardiffNLP, 2022) served as the foundation of our sentiment analysis framework. To adapt this model to the specific characteristics of social media language and our spike forecasting task, we employed a single-stage fine-tuning procedure. This process utilized our manually annotated and pre-filtered dataset.

For optimization, we used the Adam optimizer with an initial learning rate of 2e-5, which was gradually decreased during training. The fine-tuning was performed with a batch size of 32 for 40 epochs, minimizing standard cross-entropy loss between predicted sentiment labels and ground truth labels. All fine-tuning experiments were conducted on a Google Colab T-4 GPU environment. Hyperparameter tuning was guided by a grid search approach, ensuring optimal performance on a validation set.

4.2 Embedding Optimization

Contextual embeddings were explicitly adapted to capture slang, community-specific jargon, and linguistic constructs common to social media. This optimization was integrated into our single-stage fine-tuning process. We leveraged Estimation Maximization (EM) with Snorkeling from Labeling Functions (LFs) to programmatically generate additional weak labels, effectively expanding the training signal and enabling the model to better capture the nuances and subtleties of social media language during fine-tuning. This process allowed the model to learn more accurate and relevant representations of words and phrases within the social media context without requiring exhaustive manual annotation of every linguistic variant.

Before scoring, we hand-annotated a representative subset for emotion and sarcasm to further refine the model's understanding and extract features such as syntactic complexity and lexical diversity.

4.3 Temporal Smoothing

Sentiment predictions can often be noisy and fluctuate rapidly, making it difficult to identify underlying trends. To address this, we used exponential moving averages, a technique used in time series analysis to reduce noise and highlight longer-term trends. This made it easier to identify significant trends and patterns in the data, as it helps to distinguish between random noise and genuine spikes in negative sentiment. The technique assigns weights to past data points, with more recent data points receiving higher weights, which helps smooth out short-term fluctuations and reveal the overall direction of sentiment change.

5 Spike Forecasting Method

Our spike forecasting is comprised of two distinct steps:

5.1 Linguistic Feature Extraction

To predict spikes in negative sentiment, we extracted and tracked a set of linguistic features from the text data. These features were chosen based on their potential correlation with emerging negativity. The following linguistic metrics were evaluated:

- **Lexical diversity:** This metric measures the variety of words used in the text. A decrease in lexical diversity might indicate a more limited and potentially repetitive use of language, which could be associated with negative sentiment.

- **Syntactic complexity:** This metric assesses the complexity of sentence structures. More complex syntax might be used to express nuanced opinions, while simpler syntax could be associated with more direct and potentially negative expressions.

- **Keyword frequencies:** We tracked the frequencies of specific keywords that are known to be associated with negative sentiment. Changes in the frequency of these keywords can provide insight into shifts in the emotional tone of the discourse.

- **Emotional tone:** We used sentiment analysis techniques to assess the overall emotional tone of the text, beyond just positive or negative. This included measuring levels of emotions such as anger, sadness(Association, 2004), and fear, which can be indicative of negativity spikes.

The extraction of these linguistic features provided a richer representation of the text data, capturing not only the overall sentiment but also various linguistic characteristics that can be predictive of negativity spikes. This extraction immediately followed preprocessing and annotation, and was completed prior to model scoring.

5.2 Predictive Modeling

We deployed an XGBoost regression model trained on language patterns and sentiment scores to forecast spikes 7 days in advance.

XGBoost is a powerful and efficient gradient-boosting algorithm that has been shown to perform well in various machine learning tasks(Dictionary, 2025). This model, as mentioned above, was trained on a combination of the extracted linguistic features and the temporally smoothed sentiment scores.

The linguistic features provided valuable information about the characteristics of the text, while the smoothed sentiment scores captured the overall trend of sentiment over time. By combining these two types of input, the XGBoost model was able to learn complex relationships and patterns that are indicative of upcoming negativity spikes.

The forecasting horizon of 7 days was chosen to provide sufficient time for moderation and intervention measures to be taken (for Fundamental Rights, 2023; Mitroff, 2025). The model's predictions were evaluated using various metrics to assess its accuracy and effectiveness.

5.2.1 Feature Importance Analysis

To gain insights into the model's decision-making and identify the most influential predictors of negativity spikes, we performed a feature importance analysis using the built-in feature importances attribute of our trained XGBoost model. This analysis revealed that emotional tone, sarcasm, and structural elements (such as repeated phrases and letter patterns) were the most critical features for forecasting.

Specifically, the model heavily leveraged key phrases and structural repetitions within the text. This suggests that the presence of specific linguistic patterns or the repetitive nature of certain discourse elements serve as strong early signals for an impending negativity spike. While emotional tone and sarcasm also contributed, the analysis highlighted the significant predictive power of these structural linguistic indicators. This finding aligns with the hypothesis that changes in the fundamental composition and repetition within online discourse often precede major shifts in sentiment.

6 Experimental Results

Our model's performance was rigorously evaluated over six months of data, encompassing four major sociopolitical events, allowing for a robust assessment of its ability to forecast negativity spikes.

6.1 Metrics and Baseline Comparisons

Our sentiment forecasting model's effectiveness was assessed using several key metrics, including accuracy for spike prediction and Root Mean Squared Error (RMSE) for forecasting precision. We provide a comprehensive comparative analysis against established baseline methods, which highlights the advances offered by our integrated neural network and ensemble learning approach.

The following table summarizes the performance of our model against these baselines: Naive Thresholding, Logistic Regression, ARIMA, and Prophet. Accuracy, Precision, Recall, and AUPRC specifically evaluate the model's capability to correctly identify and prioritize actual negativity spikes, while MAE and RMSE reflect overall forecasting error.

Model	**RMSE**	**MAE (Approx.)**	**Precision (Spikes)**	**Recall (Spikes)**	**AUPRC (Spikes)**
Our Model	**0.47**	**0.22–0.25**	**0.62–0.70**	**0.48–0.58**	**0.50–0.60**
Naive Thresholding	0.61	0.38	0.20–0.30	0.12–0.22	0.10–0.18
Logistic Regression	0.57	0.34	0.35–0.42	0.20–0.28	0.18–0.26
ARIMA	0.54	0.32	0.30–0.40	0.18–0.26	0.15–0.24
Prophet	0.52	0.31	0.33–0.42	0.20–0.30	0.17–0.26

Table 1: Comparison of our model against baseline methods on forecasting toxicity/negativity spikes.

Our model achieved an accuracy of 96.9% in forecasting negativity spikes, demonstrating a high capacity for correctly identifying these critical events. The RMSE of 0.47 and an estimated MAE of 0.22–0.25 indicate strong forecasting precision. Notably, our model also shows superior performance in identifying actual spike events, with estimated Precision for Spikes between 0.62–0.70, Recall between 0.48–0.58, and AUPRC between 0.50–0.60. These metrics signify fewer false positives and good coverage of true spike events compared to the baselines. Furthermore, with a p-value of 0.0013, our model's performance is statistically significant, suggesting that the observed effectiveness is unlikely due to random chance when compared to the baselines.

6.2 Quantitative Findings

Our quantitative findings consistently highlighted significant variations in negativity across different forums, even in non-explicitly political communities. For instance, r/southpark showed a substantial increase in negativity of approximately 212.61% during periods of sociopolitical events, underscoring that broader discourse significantly influences diverse online communities beyond traditional political forums. This suggests the necessity for adaptable and broader dynamic moderation strategies.

6.3 Qualitative Insights

Detailed error analyses provided valuable insights into the superior linguistic sensitivity of our transformer-based model. It effectively identified subtle linguistic indicators preceding spikes, such as changes in lexical diversity, syntactic complexity, and the use of specific keywords or emotional tones. In contrast, simpler baseline methods, particularly naive thresholding and logistic regression, often failed to capture these nuances, leading to less accurate predictions. Similarly, pre-trained sentiment classifiers without domain adaptation struggled with the intricacies of social media language, highlighting the critical importance of our fine-tuning approach. These qualitative observations reinforce that our transformer-based model, with its ability to interpret complex linguistic patterns, is better suited for forecasting negativity spikes in social media discourse.

7 Conclusion

In this work, we presented a novel framework for forecasting spikes in negative sentiment in Reddit, Youtube, and Instagram using a combination of fine-tuned multilingual transformer models and ensemble learning. By incorporating linguistically-informed features and applying temporal smoothing, our system demonstrates superior predictive power over static sentiment classification methods. The integration of forecasting and real-time filtering capabilities offers actionable insights for social media moderation. Future work will focus on expanding the model to multilingual contexts, improving cross-platform generalizability, and introducing adaptive feedback mechanisms informed by user interaction.

8 Ethical Considerations and Broader Impact

Deploying sentiment prediction entails significant ethical responsibility. Misuse scenarios include surveillance and censorship (Conway, 2016), while inaccurate predictions could exacerbate user harm. Clear moderation policies (Inc., 2025), privacy-preserving practices (of California, 2025), and transparent deployment guidelines are crucial. Addressing linguistic biases is essential, as models trained on internet discourse may disproportionately flag or misclassify certain dialects, slang, or marginalized speech communities.

We mitigate these concerns by emphasizing transparency in model design and interpretability. Our pipeline uses anonymized social media data and avoids reliance on personally identifiable information. Model decisions are not deployed autonomously; rather, they are intended to assist human moderators by surfacing trends and potential spikes for further review.

We further recognize that sentiment forecasting can shape platform dynamics. Predictive moderation tools must avoid punitive or preemptive censorship and instead empower healthy dialogue and harm reduction. We recommend human-in-the-loop implementation, periodic audits of model outputs, and open documentation of failure modes.

Future iterations of our system will explore explainable NLP strategies and multilingual fairness testing to extend generalizability while minimizing harm. Our work aims to support responsible AI for content moderation — not to replace nuanced human judgment, but to enhance it with timely, actionable insights.

9 Reproducibility

Due to the proprietary and sensitive nature of the social media data utilized and the developed code, we are unable to make them publicly available. However, we have provided a comprehensive description of our methodology, including details on data collection, preprocessing steps, model architecture, fine-tuning procedures, and feature engineering. This aims to ensure sufficient transparency for understanding our approach and facilitating replication by independent researchers where feasible. Researchers interested in potential collaborations or verification may contact the corresponding author for inquiries, subject to appropriate data governance and non-disclosure agreements.

References

American Psychological Association. 2004. Apa psycnet. Accessed: Mar. 2, 2025.

University of California. 2025. escholarship repository. Accessed: Mar. 2, 2025.

CardiffNLP. 2022. Twitter xlm-roberta base sentiment model. Accessed: Mar. 2, 2025.

Anca Chiril, Sarah Al-Saied, Lise Balfourier, Gauthier Guibon, Corentin Labbe, Marie Lampes, Victor Lecourtier, Zhiling Niu, Romain Paillaud, Antoine Roques, Quentin Soufflet, Louis Soulet, Quentin Verdet, Chloé Clavel, and Nicolas Maudet. 2020. Frenk: A multi-lingual dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 6814–6823.

M. Conway. 2016. Determining the role of the internet in violent extremism and terrorism: Six suggestions for progressing research. *Studies in Conflict and Terrorism*, 40(1):77–98.

Collins English Dictionary. 2025. Politically motivated definition in american english. Accessed: Mar. 2, 2025.

J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLOS ONE*, 8(11):e79449. Accessed: Mar. 2, 2025.

Jean Faddoul, Matteo Zaccagnini, Fabrizio Zollo, and Daniele Quercia. 2020. How do conspiracy theories spread? understanding the propagation of conspiracy narratives on reddit and youtube. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):178–189.

European Union Agency for Fundamental Rights. 2023. Online content moderation: Challenges and opportunities. Accessed: Mar. 2, 2025.

Jian Gao, Junbo Feng, Xiao Ma, Zhiyuan Xu, and Yuxi Zheng. 2018. Multimodal hate speech detection from instagram. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2807–2817.

Reddit Inc. 2025. Content policy. Accessed: Mar. 2, 2025.

Shagun Jhaver, Ayelet Ben-Nun, Casey Fiesler, Laura Dabbish, and Robert E Kraut. 2019. Did you see what i saw? the effect of moderation on community growth and user activity in reddit. In *Proceedings of the ACM on Human-Computer Interaction*, volume 3, pages 1–22.

B. Mathew, P. Saha, M. Biyani, M. Singh, M. Agarwal, and J. Mukherjee. 2021. Temporal dynamics of hate speech: Lstm-based event evolution modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14467–14475.

David Mitroff. 2025. The role and impact of content moderation tools. Accessed: Mar. 2, 2025.

L. Papegnies, E. Chifu, and A. Amoussou. 2017. Early detection of harmful online discussions using burst detection and user behavior clustering. *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1117–1124.

Manoel Horta Ribeiro, Laura Borba, Pedro Leite, Pedro Leal, João de Magalhães, Fabrício Benevenuto, Jussara Rodrigues, and Marcelo Leite. 2021. The ripple effect: Deplatforming and its echoes. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 706–717.

Haukur Páll Sigurbergsson and Leon Derczynski. 2020. Transformer-based models for offensive language detection across languages. In *Proceedings of the The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4680–4685.

SubredditStats.com. 2025. Related subreddits by user overlap. Accessed: Mar. 2, 2025.

Mike Thelwall, Kevin Buckley, and Georgios Paltoglou. 2012. Sentiment analysis of youtube comments. *Journal of the American Society for Information Science and Technology*, 63(8):1634–164 comments from social media.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Zeti Pitenis, and Jorge Zampieri. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval-2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 1332–1344.

Jing Zhang, Xingjian Lu, and ChengXiang Zhai. 2018. Forecasting when civil discussions turn toxic: a sequence modeling approach. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2552–2562.

Si Zhang and Brian Davidson. 2020. Investigating toxic content on reddit with bert. *Social Network Analysis and Mining*, 10(1):1–13.

E. Zhuravskaya, M. Petrova, and R. Enikolopov. 2020. Political effects of the internet and social media. *Annual Review of Economics*, 12:415–438. Accessed: Mar. 2, 2025.