Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29656

Natural Language Processing Techniques and Methods for Identifying Negative Speech in Social Media

Shumin Wang

School of Software, Taiyuan University of Technology, Jinzhong, China
485134416@qq.com

Abstract. With the increasingly prominent role of social media in information dissemination and social interaction, its anonymity and openness have also led to the gradual prominence of negative speech, which has adverse effects on individual psychology and social stability. Traditional manual moderation models struggle to meet practical needs due to limitations like the vast volume of negative speech and the potential psychological harm to moderators. With automation, real-time processing, and scalability, Natural Language Processing (NLP) serves as a crucial tool for identifying negative language. This paper examines the concept, characteristics, and categories of negative speech on social media, offering a comprehensive analysis of the NLP frameworks employed in negative speech detection, including methods for text representation, feature extraction, and the practical applications of various models. Through a review of existing studies, this paper highlights key optimizations in various methods, identifies substantial limitations and issues in current technologies, and presents key findings on future development trends, informing research and applications in negative speech detection on social media.

Keywords: Social Media, Negative Speech Identification, Natural Language Processing (NLP), Pre-trained Models, Deep Learning

1. Introduction

As technology has advanced, social media has emerged as a key platform for the dissemination of information and the facilitation of social interaction. By the end of 2023, the global social media user base reached 4.9 billion, with users from China and India accounting for over 38%. In the first quarter of 2024 alone, Meta took action on 790 million instances of bullying and harassment and 740 million pieces of hate speech [1]. Nevertheless, the anonymity and a certain degree of openness on these platforms have reduced the constraints on user behavior, thus leading to a surge in negative speech. According to Pew Research Center data from 2018, 59% of U.S. teenagers had experienced online bullying or harassment, and during the 2020 lockdown, reported cases of cyberbullying in the U.S. increased by 56% compared to the previous year [2]. The harm caused by negative speech manifests in many ways. For individuals, exposure to such content can lead to psychological issues, including depression and anxiety, and in severe cases, even result in suicide. For society, negative speech can escalate conflicts and incite violence. Manual moderation models fail to manage the overwhelming volume of harmful speech, hence making full reliance on human review impractical.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29656

Besides, prolonged exposure to such content can lead to Post-Traumatic Stress Disorder (PTSD) in moderators, and lack of platform support has resulted in labor disputes [1]. Additionally, Natural Language Processing (NLP) technology enables automated, real-time processing at scale, making it an effective tool for identifying negative speech. Through an analysis of existing studies, this paper examines the NLP technical framework for identifying negative speech on social media, analyzes current challenges, and explores future development directions. The insights provided help to better understand the current state of the technology and inform the optimization of existing methods.

2. The concept and characteristics of negative speech on social media

2.1. Basic concept of negative speech

Negative speech refers to content on social media that is harmful or has the potential for negative impact, specifically targeting individuals or groups. It often takes the form of text, images, or videos, characterized by by aggressiveness, insults, threats, discrimination, or misinformation. Thus, such speech can lead to psychological harm, social conflict, or the spread of false information. In contrast to the general negative sentiment, which generally conveys personal emotions or subjective views, negative speech specifically targets individuals or groups with the purpose of causing harm. For example, while general negative sentiment may include expressions such as “I’m feeling sad,” negative speech involves direct insults or discriminatory remarks, such as “You’re worthless!” or discriminatory statements based on someone’s identity, leading to more significant consequences like trauma or social damage. Moreover, unlike general negative sentiment, which is often a passive emotional expression, negative speech is more likely to be actively published with harmful intent.

2.2. Classification of negative speech types

The classification of negative speech addresses the diversity and varying impacts of social media content, thereby guiding precise platform governance, distinguishing levels of harm, reducing risks, and enabling cross-regional and cross-platform comparisons. In social governance, it helps design evidence-based interventions for policymakers. By clarifying specific types of negative speech, it ensures that initiatives target the root causes rather than just surface issues. Additionally, this classification helps the public enhance media literacy; by identifying different forms of harmful content, individuals can reduce the risk of being manipulated or suffering emotional harm. Based on the content manifestations and practical impacts of negative speech, it can generally be divided into three categories: Emotional, Aggressive, and Informational, as shown in Figure 1.

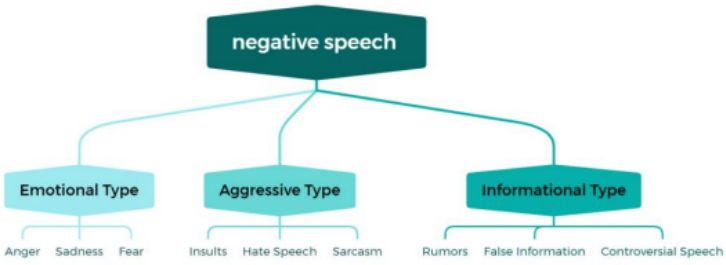

Figure 1. Classification chart of negative remarks

More specifically, emotional negative speech shows intense negative emotions, thus triggering resonance or psychological discomfort in others and is categorized by emotional tone, with anger

DOI: 10.54254/2755-2721/2026.TJ29656

speech highlighting expression and blame, sadness speech conveying despair and sometimes amplifying emotions or spreading harmful ideas, and fear speech seeking to intimidate, create chaos, or attract attention. Besides, aggressive negative speech directly attacks or insults individuals or groups, representing the most immediately harmful type, and is further divided by attack form into insulting speech, which uses vulgar or offensive language for direct attacks, hate speech, which targets individuals or groups according to identity characteristics such as race, ethnicity, or religion, and sarcastic or mocking speech, which harms feelings or reputations through indirect means like innuendo. Moreover, informational negative speech misleads with false or biased information and is categorized by information type, including rumors that lack factual basis and are often offensive, false information or fake news that is fabricated, altered, or pieced together and inconsistent with facts, and controversial or biased speech that presents highly biased content on specific topics or targets, provoking public disputes or conflicts.

2.3. Social media expression characteristics

The expression of negative speech on social media is typically shaped by platform characteristics and user habits. Different platforms foster unique user expression habits and content forms, which in turn significantly increase the difficulty of identifying negative speech through NLP technology.

From the perspective of textual features, various platforms impose character limits. For example, Twitter/X has a 280-character limit and Weibo has a 140-character limit, so negative speech often appears in the form of short texts. However, this conciseness leads to ambiguity in expression, and additional context is sometimes required to determine whether a statement constitutes an attack. Beyond short texts, users also widely use colloquial expressions, abbreviations, and internet slang, such as the English colloquial term “gonna,” the Chinese homophonic phrase “绝绝子” (meaning extremely excellent), and the abbreviation “YYDS” (which means eternally the best or top-tier). And traditional text processing methods struggle to handle these elements, which increases the difficulty of identification. At the same time, users may intentionally alter the spelling of offensive words to evade detection, further raising the challenge for NLP techniques in feature matching.

With the continuous development of social platform functionalities, negative speech is no longer limited to pure text but has begun to integrate multimodal information such as images, videos, and emojis. Common approaches include combining images with text, pairing videos with text, and using emojis to supplement expression; all these methods have expanded the forms of negative speech expression. Besides, due to differences in user bases and functional designs across platforms, platform variations also result in distinct forms of expression and distribution patterns for negative speech: Twitter/X primarily relies on short texts and hashtags, with a relatively high proportion of hate speech and politically controversial content; Facebook and Instagram have a higher proportion of image and video content, where cyberbullying and personal attack-related negative speech are more common; on Twitch, negative speech mainly appears in bullet comments within live chat, and such speech is highly real-time, mostly consisting of insulting and provocative content.

3. Methods and applications for identifying negative speech on social media

3.1. Text representation and feature construction

In text analysis and negative speech detection, he evolution of feature representation methods has always been the core driving force for advancing technological development, and its development

DOI: 10.54254/2755-2721/2026.TJ29656

process can be divided into three main stages: traditional features, word vector representations, and contextual embeddings.

Traditional feature methods are represented by the Bag-of-Words (BoW) model and sentiment lexicons, relying on rules or statistical approaches to construct features and performing well in early small-scale data scenarios. The BoW model treats text as an unordered set of words, counting word frequency while ignoring word order and semantic information; for example, “I hate you” and “You hate me” receive the same representation [3]. Sentiment lexicons label the emotional tendency of text, and scholars such as Fortunatus calculated a “text aggression score” by integrating multiple lexicons to assist in detecting cyberbullying content [2]. However, these methods are limited when encountering internet slang or emerging vocabulary.

Word vector representation methods learn low-dimensional dense vectors of words via neural networks, capturing semantic relationships and bridging the “semantic gap” in traditional features. Mainstream methods include Word2Vec and GloVe. The CBOW model in Word2Vec predicts a central word from its context, while GloVe combines global co-occurrence statistics and local context to construct word vectors. In cross-language negative speech detection, GloVe outperforms Word2Vec but has limited capability in handling out-of-vocabulary (OOV) words. By building on word vector approaches, contextual embedding methods employ Transformer-based pre-trained models to learn contextual dependencies from large-scale corpora, effectively resolving polysemy and semantic ambiguity. Mainstream models include Bidirectional Encoder Representations from Transformers (BERT) and HateBERT. Specifically, BERT uses a bidirectional Transformer and is pre-trained with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), hence dynamically adjusting word representations based on context. In contrast, HateBERT is retrained on a Reddit offensive speech dataset to boost recognition of negative texts like hate speech, achieving higher F1 scores than general BERT in hate speech detection tasks and exhibiting better scenario adaptability, as shown in Figure 2 [4,5].

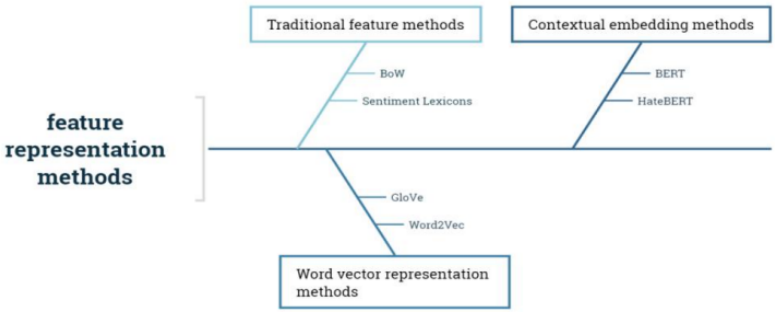

Figure 2. Feature representation methods

3.2. Model methods and application practices

In the field of negative speech detection on social media, traditional manual rule-based methods and shallow machine learning approaches have gradually become insufficient due to their inability to adapt to text characteristics such as implicit semantics, irregular expression, and strong context dependence. In contrast, neural networks have become the main approach, using deep structures and self-learning. Specifically, they extract semantic features automatically via multi-layer networks, removing the need for manual feature engineering, and handle polysemy and contextual ambiguity by modeling long-range dependencies with Recurrent Neural Networks (RNNs) and Transformers. Additionally, neural networks exhibit strong scenario adaptability, as they can be trained for domain adaptation to accommodate stylistic differences in text across various platforms.

DOI: 10.54254/2755-2721/2026.TJ29656

In practical applications, BERT and its variants optimized the understanding of long texts using bidirectional encoding. More recently, a double-layer hybrid CNN-RNN model proposed by Riyadi et al. achieved a significant performance breakthrough, employing Convolutional Neural Networks (CNNs) to extract local semantic patterns and Long Short-Term Memory (LSTM) networks to capture long-range dependencies. On imbalanced Twitter hate speech datasets, the model reached an F1-score of 0.883, representing a 34.4 percent improvement over traditional single-layer hybrid models. After balancing the data, its F1-score further increased to 0.914 [6]. Moreover, the effective use of Dropout and early stopping techniques mitigated overfitting, fully verifying the accuracy and stability of neural networks in negative speech detection.

3.3. Method comparison and optimization strategies

In negative speech recognition, different methods vary considerably in their strengths, weaknesses, and performance across cross-platform and cross-lingual scenarios. Traditional methods are simple and computationally efficient but rely heavily on platform-specific features such as hashtags and emoticons, thus leading to poor cross-platform generalization. Besides, they lack adaptability in cross-lingual contexts, which limits their effectiveness in multilingual recognition. Monolingual pre-trained models, such as BERT, capture universal language representations and achieve strong cross-platform performance. As noted by Paul, tests on Twitter, Wikipedia, and Formspring showed that BERT achieved an average F1-score of 81%, outperforming CNN (72%) and SVM (68%) [7]. However, these models perform well only in single-language scenarios and require large amounts of training data. Multilingual pre-trained models address these cross-lingual limitations. For example, Kumaresan et al. applied such models to detect homophobic and transphobic speech in low-resource Dravidian languages, achieving an F1-score of 73%, 21% higher than English BERT [8]. However, due to limited data in low-resource languages, their performance remains below that in English scenarios (85%), and their adaptation to platform-specific features is weaker than that of models optimized for specific platforms.

To address the shortcomings of the aforementioned methods, specific optimization efforts can be implemented from multiple dimensions to improve the effectiveness of negative speech recognition. In terms of multimodal fusion, a hybrid early and late fusion strategy is adopted. Initially, text and emoji vectors are fused to incorporate semantic and emotional information, after which attention mechanisms assign weights to image features to emphasize critical visual cues. And this approach improves multimodal integration and mitigates the incomplete representation inherent in text-only features. In transfer learning for data-scarce scenarios, tthe model is pre-trained on source data, adjusted with Maximum Mean Discrepancy (MMD) to reduce domain differences, and fine-tuned on the target domain. For example, the MMD-LCDH model pre-trained with Llama2 applied MMD to reduce domain distribution differences, improving cross-domain hate speech F1-score by 18% [9]. This effectively solves the problem of poor model generalization caused by insufficient data. At the level of pre-trained model adaptation, pre-training tasks are customized for negative speech scenarios, such as hate speech masked prediction and aggressive intent classification, allowing the model to focus more on learning negative semantic features. By enhancing feature engineering, user features are combined with text features to further enrich the information dimension, improve the accuracy of cyberbullying detection, and enhance the model's ability to recognize negative speech associated with user behavior.

DOI: 10.54254/2755-2721/2026.TJ29656

4. Existing problems and future efforts

Language styles and expression habits vary significantly across different platforms. When models are applied to a target domain different from the training data source, their performance often drops sharply. Moreover, the semantic gap in cross-lingual transfer further exacerbates performance loss. Multimodal fusion provides a new perspective to tackle this issue, surpassing traditional feature concatenation models. By utilizing modality attention and cross-modal Transformers, it aligns text, image, and video semantics, and incorporates commonsense and domain knowledge to construct negative speech knowledge graphs, enhancing semantic clarity and context comprehension.

Few-shot learning and cross-domain adaptation pose significant difficulties. Limited labeled data in low-resource languages and specialized platforms impede effective feature learning, while pronounced variations in language styles and expression habits across platforms lead to substantial performance drops when models are transferred to target domains differing from the training source. To address this issue, a “pre-training plus prompt engineering” approach is employed to boost model performance in low-data scenarios. Self-supervised learning generates pseudo-labels from unlabeled data, reducing dependence on manual annotation, while reinforcement learning uses identification accuracy and generalization as reward metrics to dynamically optimize model parameters and enhance overall performance.

From a social and ethical perspective, technology will increasingly prioritize privacy, fairness, and rational regulation. Meanwhile, the over-representation of negative samples targeting a specific group can lead models to produce discriminatory outputs, undermining the fairness of identification results. For the group biases existing in training data, models will be optimized via bias detection and fairness constraint methods to reduce discriminatory outputs. Besides, social impact assessment metrics for negative speech detection will be established to avoid over-detection and ensure the appropriateness and rationality of regulatory measures.

5. Conclusion

This study reveals that the identification of negative speech on social media plays a crucial role in maintaining a healthy online environment, protecting individual rights, and promoting social harmony. With the iterative advancement of NLP technologies, this task has become more accurate and efficient. Traditional machine learning methods, which rely on manually designed features such as TF-IDF and n-grams, have surpassed the limitations of early keyword-matching approaches but remain limited in capturing deep semantics and handling context-dependent speech. The adoption of deep learning significantly addressed these shortcomings. Pre-trained models such as BERT and RoBERTa, trained on large-scale corpora to acquire universal semantic knowledge and fine-tuned for specific tasks, can deeply understand contextual differences and linguistic nuances, showing stronger cross-domain generalization than traditional methods. Multimodal fusion combines speech features, such as intonation and pace, with text semantics to overcome dialect misclassification and semantic ambiguity. The evolution of NLP technologies, from feature engineering to autonomous semantic learning and from single-modal to multimodal analysis, continues to enhance the accuracy, generalization, and adaptability of negative speech detection.

References

[1] Saleous, H., Gergely, M., & Shuaib, K. (2025). Exploring NLP-based solutions to social media moderation challenges. Wiley Human Behavior and Emerging Technologies, 9436490..

DOI: 10.54254/2755-2721/2026.TJ29656

[2] Al-Harigy, L. M., Al-Nuaim, H. A., Moradpoor, N., & Tan, Z. (2022). Building towards automated cyberbullying detection: A comparative analysis. Hindawi Computational Intelligence and Neuroscience, 4794227.
[3] Dogra, V., Verma, S., Kavita, P., Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A complete process of text classification system using state-of-the-art NLP models. Computational Intelligence and Neuroscience, 1883698.
[4] Liu, P., Li, W., & Zou, L. (2019). NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. Proceedings of SemEval-2019 Task 6, 87-91.
[5] Caselli, T., Basive, V., Mitrovic, J., et al. (2020). HateBERT: Retraining BERT for abusive language detection in English, 4-5.
[6] Riyadi, S., Andriyani, A. D. Y., & Sulaiman, S. N. (2024). Improving hate speech detection using double-layer hybrid CNN-RNN model on imbalanced dataset. IEEE Access, 12, 159660-159668.
[7] Yadav, U., Bondre, S., Thakre, B., et al. (2024). Speech-to-text Emotion Detection System using SVM, CNN, and BERT. In 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE), 1-5.
[8] Sharma, D., Gupta, V., & Singh, V. (2023). Detection of homophobia & transphobia in Dravidian languages: Exploring deep learning methods. In International Conference on Advanced Network Technologies and Intelligent Computing, 1798, 225-236.
[9] Li, Z.Q. (2023). A Study on Cross-Domain Multimodal Hate Speech Detection. Dalian University of Technology.