International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

Emotional Tone Detection in Hate Speech Using Machine Learning and NLP: Methods, Challenges, and Future Directions, a Systematic Review.

INPLASY2025110006
doi: 10.37766/inplasy2025.11.0006
Received: 3 November 2025
Published: 4 November 2025

Escobar, A; Rivadeneira, R; Fuertes, W.

Corresponding author:
Walter Fuertes

wmfuertes@espe.edu.ec

Author Affiliation:
Universidad de las Fuerzas Armadas ESPE.

ADMINISTRATIVE INFORMATION

Support - Universidad de las Fuerzas Armadas ESPE.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY2025110006

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 4 November 2025 and was last updated on 4 November 2025.

INTRODUCTION

Review question / Objective This research aims to identify hate speech on social networks using techniques based on Natural Language Processing (NLP) and machine learning (ML), considering emotional tone as a component to improve the accuracy of detection models.
The research questions were as follows:
RQ1: What are the most commonly used NLP and machine learning techniques to detect emotional tone and/or hate speech on social networks?
RQ2: What tools do authors use to implement NLP and machine learning techniques to detect hate speech and/or emotional tone on social media?
RQ3: What are the most detected emotions in hate speech?
RQ4: What are the main challenges, limitations, and future research directions for using NLP and machine learning techniques to detect emotional tone in hate speech?
RQ5: Which NLP or machine learning models perform best in the classification of emotional tone based on metrics such as precision, recall, or the F1 score?

Condition being studied The use of social networks has transformed the ways we communicate, interact, and disseminate opinions. However, these environments facilitate the spread of hate speech and cyberbullying, which most severely impact vulnerable groups. This form of digital violence generates various psychological, social, and emotional effects, making the automated detection of this discourse a research challenge.

METHODS

Search strategy We applied the PICOS (Population, Intervention, Comparison, Outcome, and Study Design) method to develop our search strategy, allowing for more precise and targeted search strategies. Also, this achieves greater

precision in retrieving relevant studies by simplifying the identification of studies that meet the established criteria in each category. Therefore, PICOS improves the process of selecting and analyzing available scientific evidence.
P: “Hate Speech”, “Online Hate Speech”, “Hate Speech Against Women”, “Offensive Messages on Social Media”, “Hate Messages Against Women”, “Emotional tone”
I: “Natural Language Processing”, “Machine Learning”, “Techniques”, “Classification”, “Supervised/Unsupervised Machine Learning”, “RNN”, “BERT”, “GPT”, “Emotion Detection”
C: “Deep Learning Models and Pre-Trained Embeddings vs. Traditional NLP Classification Techniques”
O: “Precision”, “Accuracy”, “Recall”, “Detection Rate”, “Precision”, “F1-Score”, “False Positive Rate”, “False Negative Rate”.
S: “Empirical Studies”, “Comparative Analyses”, “Correlational Studies”, “Inferential Statistical Analysis”.
Terms and electronic databases included in the review.
IEEE Xplore & ("hate speech" OR "hate speech against women") AND "emotion detection" AND ("natural language processing" OR "machine learning" OR BERT OR GPT OR "deep learning")
Science Direct & ("hate speech" OR "hate speech against women") AND "emotion detection" AND ("natural language processing" OR "machine learning" OR BERT OR GPT OR "deep learning").
ACM & ("hate speech" OR "hate speech against women") AND "emotion detection" AND ("natural language processing" OR "machine learning" OR BERT OR GPT OR "deep learning").
Springer & ("hate speech" OR "hate speech against women") AND "emotion detection" AND ("natural language processing" OR "machine learning" OR BERT OR GPT OR "deep learning").
WILEY & ("hate speech" OR "hate speech against women") AND "emotion detection" AND ("natural language processing" OR "machine learning" OR BERT OR GPT OR "deep learning").

Participant or population This systematic review will address primary studies that involve the participation of research teams or systems applying Natural Language Processing (NLP) and Machine Learning (ML) techniques to detect emotional tone in hate speech. The “participants” in the context of this review are not human subjects, but rather published studies presenting quantitative analyses and measurable outcomes related to hate speech detection models. Only peer-reviewed studies written in English and published between 2019 and 2025 will be included.

Intervention The interventions evaluated in this review correspond to computational approaches and methodological frameworks based on Natural Language Processing (NLP) and Machine Learning (ML) techniques applied to the detection of emotional tone in hate speech. These interventions include model architectures, feature extraction methods, linguistic preprocessing techniques, and algorithmic strategies aimed at improving the accuracy and reliability of hate speech classification.

Comparator Comparative interventions in this review include alternative NLP and ML techniques or model configurations used to detect hate speech and analyze emotional tone. These may involve comparisons among different algorithmic approaches (e.g., traditional machine learning models versus deep learning architectures), distinct feature extraction methods (lexical, semantic, or contextual embeddings), or various preprocessing strategies. The purpose of these comparisons is to identify which methodological approaches achieve superior quantitative performance in detecting emotional tone within hate speech content.

Study designs to be included The review will include quantitative primary studies that present experimental or comparative research designs related to the detection of emotional tone in hate speech using NLP and ML techniques. Eligible study designs include empirical evaluations, benchmark experiments, cross-validation studies, and comparative analyses of model performance. Only peer-reviewed publications that report measurable outcomes such as accuracy, precision, recall, F1-score, or AUC will be included. Qualitative, theoretical, or purely descriptive works will be excluded.

Eligibility criteria We collected data from selected primary articles related to the natural language processing and machine learning techniques employed. We focused on the validation methods and techniques used, the datasets applied, the software artifacts implemented, and the evaluation metrics for emotional tone classification in hate speech. We defined the inclusion and exclusion criteria described below.

Primary studies use NLP and ML techniques to detect emotional tone in hate speech.
Only studies with quantitative results, i.e., with precise measurements.
Written only in English.
From the last 6 years (2019 – 2025).

Exclusion Criteria
Studies that do not present empirical results related to the detection of emotional tone in hate speech.
Research that uses datasets that are unrepresentative, irrelevant, or unrelated to hate speech.
Studies that lack a precise, reproducible, and evaluable methodology for emotional tone classification.
Studies focused on theoretical or conceptual aspects of natural language processing or machine learning without providing applicable or measurable results.

Information sources The information sources for this systematic review include major academic databases and digital libraries relevant to computer science and artificial intelligence. We utilized IEEE Xplore, ScienceDirect, ACM Digital Library, SpringerLink, and Wiley Online Library as the primary sources of indexed peer-reviewed articles. These databases were selected for their comprehensive coverage of research on Natural Language Processing (NLP), Machine Learning (ML), and computational linguistics applied to hate speech detection.

Main outcome(s) The main outcomes of this systematic review focus on the effectiveness of Natural Language Processing (NLP) and Machine Learning (ML) techniques in detecting emotional tone within hate speech content. Across the 34 primary studies analyzed, the results demonstrate that these computational approaches enable the identification of emotional patterns directed toward vulnerable or marginalized groups.
Among the most frequently employed techniques are those emphasizing linguistic preprocessing and semantic embeddings, particularly Word2Vec and BERT. The review revealed a consistent preference for models such as BERT, RoBERTa, Support Vector Machines (SVM), and Random Forests, which exhibit robust performance across various datasets. Notably, some less common architectures, such as LLaMA 2, demonstrated superior results in specific contexts, suggesting that emerging large language models hold strong potential for future studies.
The emotional patterns identified show a predominance of negative emotions—notably anger and fear—aligned with the inherently hostile characteristics of hate speech. Additionally, the ironic or sarcastic use of positive emotions was detected, which introduces semantic ambiguity and presents significant challenges for automated classification systems.
The effect measures reported across studies include quantitative performance metrics such as accuracy, precision, recall, F1-score, and AUC, allowing for comparative evaluation of model efficiency. The main limitations observed relate to the scarcity of multilingual datasets, the limited cultural diversity represented in training corpora, and the difficulty of interpreting nuanced expressions such as sarcasm and irony. These findings highlight the need for more comprehensive, linguistically inclusive resources and hybrid modeling strategies for emotion-aware hate speech detection.

Would you like me to help you now with the next INPLASY section — Search strategy, describing how you built and combined your keywords and Boolean operators?

Quality assessment / Risk of bias analysis The quality assessment of the primary studies followed the methodological guidelines of the PRISMA 2020 Statement and the PICOS framework (Population, Intervention, Comparison, Outcomes, and Study Design). The process was designed to ensure methodological rigor, reproducibility, and analytical depth throughout the review.
A statistical processing phase was conducted to normalize performance metrics and identify central tendencies and dispersion across models. Descriptive statistics and comparative tables were used to evaluate convergence or divergence among studies. Analytical and synthetic capacity was applied to integrate these quantitative outcomes into coherent insights, highlighting methodological patterns, common biases, and areas of improvement.

Strategy of data synthesis The data synthesis strategy was designed to integrate quantitative evidence from the selected studies and identify methodological trends, performance variations, and emerging research patterns in the application of NLP and ML techniques for emotional tone detection in hate speech.

A quantitative descriptive synthesis was conducted to summarize the statistical results reported by each study, including performance metrics such as accuracy, precision, recall, F1-score, and AUC. When possible, measures of central tendency (mean, median) and dispersion (standard deviation, range) were calculated to facilitate comparison among different models and techniques. The synthesis emphasized the relative performance of traditional machine learning algorithms (e.g., SVM, Random Forest) versus

deep learning models (e.g., BERT, RoBERTa, LLaMA 2).

The analytical process followed the PRISMA 2020 guidelines, applying an iterative approach of data extraction, verification, and coding. Data were tabulated and categorized according to model type, feature representation, dataset language, and emotion classification strategy. This enabled cross-comparison and clustering of methodological approaches.

Where statistical aggregation was not feasible due to heterogeneity in datasets or evaluation protocols, a narrative synthesis was used to interpret trends and highlight consistent patterns in quantitative findings. The analysis also incorporated statistical visualization techniques (e.g., comparative tables, frequency plots) to enhance interpretability.

Subgroup analysis Subgroup analyses were conducted to explore variations in model performance and methodological outcomes across specific dimensions relevant to the detection of emotional tone in hate speech using NLP and ML techniques. These analyses aimed to identify how contextual, linguistic, and algorithmic factors influence quantitative results and model robustness.
The subgroups were defined according to the following criteria:
1. Model Type: Comparison between traditional machine learning algorithms (e.g., SVM, Random Forest) and deep learning architectures (e.g., BERT, RoBERTa, LLaMA 2).
2. Feature Representation: Distinction between studies employing lexical features (e.g., bag-of-words, TF-IDF) versus semantic or contextual embeddings (e.g., Word2Vec, GloVe, BERT embeddings).
3. Language of Dataset: Analysis of models trained on English corpora versus those incorporating multilingual or cross-lingual datasets.
4. Emotion Category: Examination of model performance according to specific emotional tones detected (anger, fear, disgust, or irony).
5. Publication Year: Assessment of potential temporal evolution or improvement in model accuracy between 2019 and 2025.
Each subgroup comparison involved descriptive statistical analysis of reported metrics (accuracy, precision, recall, F1-score, and AUC) to determine consistent performance patterns. When feasible, mean differences and standard deviations were calculated to evaluate performance variability.

Sensitivity analysis A sensitivity analysis was performed to evaluate the robustness and reliability of the synthesized findings regarding the effectiveness of NLP and ML techniques in detecting emotional tone in hate speech. This process aimed to determine whether the overall conclusions of the review were influenced by methodological decisions, data variability, or the inclusion of specific studies.
Statistical recalculations and descriptive comparisons were used to measure deviation in mean performance values and standard deviations across these subsets. Minimal variation among subgroup results was interpreted as evidence of consistency and reliability, whereas larger deviations indicated potential model or dataset-dependent effects.

Overall, the sensitivity analysis strengthened the interpretive validity of the review by confirming that the synthesized conclusions were not disproportionately affected by individual studies or methodological heterogeneity, ensuring a more stable and generalizable understanding of the emotional-tone detection capabilities in hate-speech analysis.
The analysis involved reassessing aggregated performance metrics, including accuracy, precision, recall, F1-score, and AUC after selectively excluding studies identified as having moderate or low methodological quality based on the PRISMA-PICOS assessment criteria. The goal was to verify whether the exclusion of these studies produced significant changes in the central tendencies or ranking of model performance.

Country(ies) involved The study is being carried out in Ecuador, with a global scope including primary studies conducted worldwide. The authors work at the Universidad de las Fuerzas Armadas ESPE.

Keywords Emotional tone; hate speech; machine learning; NLP; PICOS; PRISMA; SLR.

Contributions of each author
Author 1 - Aymé Escobar Díaz - Conceptualization, methodology, literature search, data curation, and original draft preparation.
Email: aaescobar2@espe.edu.ec
Author 2 - Ricardo Rivadeneira - Conceptualization, methodology, formal analysis, validation, and visualization.
Email: rxrivadeneira1@espe.edu.ec
Author 3 - Walter Fuertes - Supervision, project administration, validation, formal analysis, and review and editing of the manuscript.
Email: wmfuertes@espe.edu.ec