Review

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

Aymé Escobar Díaz , Ricardo Rivadeneira , Walter Fuertes *

Posted Date: 31 October 2025

doi: 10.20944/preprints202510.2435.v1

Keywords: emotional tone; hate speech; machine learning; NLP; PICOS; PRISMA; SLR

 doi:10.20944/preprints202510.2435.v1

Review

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

Aymé Escobar Díaz, Ricardo Rivadeneira and Walter Fuertes *

Department of Computer Science, Universidad de las Fuerzas Armadas (ESPE), Av. General Rumiñahui 171103, Quito, Ecuador
* Correspondence: wmfuertes@espe.edu.ec

Abstract

Hate speech is a form of communicative expression that promotes or incites unjustified violence. The increase in hate speech on social media has prompted the development of automated tools for its detection, especially those that integrate emotional tone analysis. This study presents a systematic review of the literature, employing a combination of PRISMA and PICOS methodologies to identify the most used Machine Learning techniques and Natural Language Processing emotion classification in hostile messages. It also seeks to determine which models and tools predominate in the analyzed studies. The findings highlight LLaMA 2 and HingRoBERTa, achieving F1 scores of 100% and 98.45%, respectively. Furthermore, key challenges are identified, including linguistic bias, language ambiguity, and the high computational demands of some models. This review contributes an updated overview of the state of the art, highlighting the need for more inclusive, efficient, and interpretable approaches to improve automated moderation on digital platforms. Additionally, include techniques, methods, and future directions in this topic.

Keywords: emotional tone; hate speech; machine learning; NLP; PICOS; PRISMA; SLR

1. Introduction

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.

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.

NLP also addresses situations related to offensive language in digital environments. In particular, the use of transformers marked a breakthrough in the accuracy of automatic text classification systems [1]. However, traditional approaches based on sentiment polarity are insufficient to capture the emotional complexity present in hostile messages.

On the other hand, emotion analysis has become established as an alternative that allows us to identify whether a message conveys a negative emotion, as well as to distinguish the type of emotion expressed, such as anger, fear, or sadness [2]. This capability is valuable when hate speech is expressed indirectly or ambiguously, making it difficult to detect using conventional methods [3]. Therefore, systems that incorporate emotion analysis can obtain greater sensitivity to implicit hostility patterns.

Furthermore, the use of a multimodal architecture enables the integration of contextual, linguistic, and behavioral cues, thus improving the accuracy of tasks such as the detection of cyberbullying [4]. Furthermore, recent studies have indicated that emotion classification also expands the models' ability to adapt to texts in minority languages or with dialectal variation [5].

 doi:10.20944/preprints202510.2435.v1

Another aspect to consider is the application of multilingual and multi-technique models that analyze content in different cultural and linguistic contexts, which is key to addressing the diversity of hate speech directed at women on a global scale [6]. Furthermore, the use of Explainable Artificial Intelligence (XAI) techniques gained relevance by providing transparency in the models' decisions and strengthening their reliability in sensitive contexts [7].

To organize and systematize recent research on this topic, we conducted a structured review using the PRISMA methodology [8]. The PICOS model was also used to formulate research questions precisely, delimiting the population as social media users, the intervention as the use of NLP, ML, emotional analysis, the comparison between traditional and advanced models, and the result in terms of system precision [9].

The results show the most widely used techniques for classifying emotions and hate speech, the methods employed by researchers, the most commonly recognized emotions, the main methodological challenges, the opportunities and gaps identified in the use of these techniques, the models that report the best performance according to metrics such as precision, recall, and F1-score, as well as lines of future research. This information is presented in a structured manner that facilitates an understanding of the current state of the field, its advances, limitations, and emerging trends.

Among the contributions of this study are: i) the characterization of the NLP techniques and ML models most used in the classification of emotional tone in hate speech, based on the analysis of their frequency and the role they play in the textual analysis process; ii) the identification of the most frequently used methods and algorithms and the models that report the best results according to the evaluation metrics for classification models; iii) an updated overview available to researchers on existing strategies for emotional and hate speech detection, facilitating comparison between approaches and highlighting those that show the most excellent effectiveness; iv) a list of the challenges, opportunities, and lines of future research, among which the integration of the emotional dimension, the improvement in tone moderation in hate speech on social networks, and the definition of adequate criteria for the selection and training of artificial intelligence models, as a fundamental part in the construction of software artifacts oriented to its prevention.

The remainder of the article is structured as follows. Section 2 details the procedure used for bibliographical research, combining the PRISMA and PICOS methodologies. Section 3 presents the evaluation of the results and their discussion, highlighting the 34 primary studies obtained and performing analysis and interpretation for each research question. Finally, Section 5 presents the conclusions and lists future lines of work.

2. Materials and Methods

This section outlines the methodological process employed in conducting the systematic review of the literature, which was based on the principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). PRISMA entails a methodical strategy that aims to conduct systematic reviews, ensuring transparency and methodological rigor throughout the review process [1]. PRISMA establishes a set of criteria for the research process, from the collection and selection of articles to the analysis and interpretation of results [10]. Figure 1 illustrates the flowchart outlining the selection of studies and the application of the checklist, which ensures a clear and comprehensive presentation of methods and results [8]. It also shows the total number of records processed, the criteria applied in each phase, and the final number of primary studies selected. Within this methodological approach, we conducted a critical evaluation of the selected studies, considering key elements such as study design, data quality, the assessment of classification model metrics (i.e., precision, recall, F1-score), and the clarity of the description of the NLP techniques and machine learning models employed [11,12]. This assessment enabled us to select relevant studies, considering the quality of their content and results.

 doi:10.20944/preprints202510.2435.v1

2.1. Research Questions

Given that this study focuses on selecting the machine learning (ML) and natural language processing (NLP) methods and techniques used to detect the emotional tone of hate speech against women present on social media, which perpetuates inequality and gender stereotypes, it is crucial to identify where the line between current knowledge and ignorance lies [13]. Below, we establish the relevant research questions.

- **RQ1:** What are the most used NLP and machine learning techniques for detecting emotional tone and/or hate speech on social media?
- **RQ2:** What tools do authors use to implement NLP and machine learning techniques for detecting 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 for detecting emotional tone in hate speech?
- **RQ5:** Which NLP or machine learning models perform best in emotional tone classification based on metrics such as precision, recall, or F1 score?

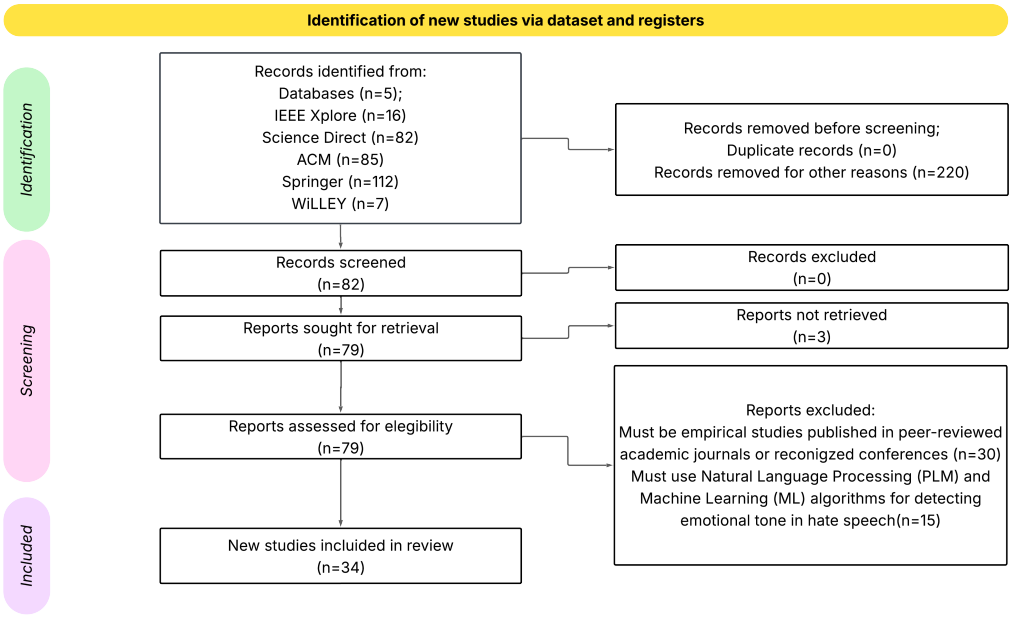

Figure 1. The PRISMA 2020 Flowchart illustrates the procedure and preliminary results of this research. Source: Adapted from [8] .

2.2. Eligibility Criteria

We only included studies that used Natural Language Processing and Machine Learning techniques to detect emotional tone in hate speech.

2.3. Data Sources

We utilized IEEE Xplore, ScienceDirect, ACM Digital Library, Springer, and Wiley, which are digital databases that provide access to relevant research in engineering and computer science.

2.4. 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 [9]. Table 1 lists the components of the method:

 doi:10.20944/preprints202510.2435.v1

Table 1. Components of PICOS.

Component	Definition	Guiding Question
P	**Population**	Who am I referring to?
I	**Intervention**	What action or intervention do I want to study?
C	**Comparison**	What can I compare it to? Are there other options?
O	**Outcomes**	What effect or outcome do I hope to observe?
S	**Study Design**	What type of study will I use to answer the question?

- **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”.

We then defined the search strings and their variants using PICOS, combining terms with Boolean operators to create a comprehensive search. We applied the following search strings to each digital database. Table 2 presents the number of articles found.

Table 2. Articles found according to the search strings applied.

Database	Search Strings	Number of Articles
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")	16
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")	82
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")	85
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")	112
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")	7
Total		**302**

We conducted initial searches in April 2025 using the databases listed above, which were selected for their availability and access to scientific literature. We subsequently expanded the search strategy by identifying keywords related to the PICOS method, combined with the Boolean operators "AND" and "OR" as described in Table 2, yielding a total of 302 articles. Before proceeding with the selection of articles. We defined the inclusion and exclusion criteria described below.

 doi:10.20944/preprints202510.2435.v1

2.5. Inclusion Criteria

- Primary studies using 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).

2.6. 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.

2.7. Data Classification

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.

3. Results

This section presents the results obtained. In summary, we identified 82 relevant articles compiled from scientific databases, including IEEE Xplore, ScienceDirect, ACM, Springer, and Wiley, as shown in Figure 1. After applying the inclusion and exclusion criteria, we eliminated 48 articles that did not focus on the classification of emotional tone in hate speech using NPL and ML techniques, were not empirical studies, and did not present quantitative results. Therefore, we selected 34 primary articles for analysis. Subsequently, each article was read, reviewed, and studied to answer the research questions individually and classify the collected information. We present the analysis of the results below, grouped by each research question:

RQ1: *What are the most used NLP and machine learning techniques for detecting emotional tone and/or hate speech on social media?*

We identified several NLP techniques for detecting emotional tone in hate speech. Their respective references are listed in Table 3, sorted in descending order by number of articles (i.e., “frequency”). We also include a "category" column, which indicates the function of each technique within the language processing pipeline. This classification distinguishes whether the method corresponds to text preprocessing, vector representation, feature extraction, and semantic or emotional analysis.

Table 3 lists the most common techniques. The most used method is tokenization (17 studies), classified as a part of textual preprocessing, which is followed by a bag-of-words (11 studies) approach, a classic textual representation technique. Within semantic and emotional analysis, emotional processing is particularly notable (9 studies). Approaches linked to feature extraction and modern contextual representation, such as lexical features and BERT embeddings (each with eight studies), are also common. Other relevant techniques include stopword removal (in 6 studies) and vector representations, such as Word2Vec, GloVe, and FastText (each with five studies). In addition, linguistic normalization methods are identified, such as lemmatization and contextual embeddings (4 studies). Additionally, we find other less frequently used techniques such as stemming, feature engineering, and attention mechanisms (3 studies).

 doi:10.20944/preprints202510.2435.v1

Table 3. Techniques using NLP are found in SLR.

Technique	Articles/Authors	Category	Frequency
Tokenization	[3,4,14–29]	Text preprocessing	17
Bag of Words	[3,5,17,19–21,25,27,30–32]	Traditional text representation	11
Emotional Processing	[7,14,18,22,25,31,33,34]	Semantic/emotional analysis	9
Lexical Features	[7,14,18,31,34–36]	Feature extraction	8
BERT Embeddings	[3–5,14–16,20,37]	Modern contextual representation	8
Stopword Removal	[17,20,21,25,27,30]	Text preprocessing	6
Word2Vec	[5,22,30,31,38]	Vector representation	5
GloVe	[3,19,30,31,39]	Vector representation	5
FastText	[5,15,19,27,39]	Vector representation	5
Lemmatization	[14,27,30,39]	Linguistic normalization	4
Contextual Embeddings	[3,5,14,22]	Modern contextual representation	4
Stemming	[17,25,39]	Linguistic normalization	3
Feature Engineering	[25,26,40]	Feature extraction	3
Attention Mechanisms	[7,24,38]	Interpretation mechanisms	3
Total			**91**

The frequency of use of NLP techniques reveals the most widely used tools and, at the same time, indicates changes in the way the field is evolving. We find the simultaneous presence of traditional approaches, such as the bag-of-words model, alongside recent models, including BERT and attention mechanisms. The results demonstrate that classical techniques play a functional role due to their low computational cost and ease of implementation in diverse environments. At the same time, the increasing use of contextual embedding reflects a trend toward models that take into account the context in which words are used, which is key to identifying the emotional tone in hate speech.

On the other hand, the low frequency of methods such as stemming, feature engineering, and attention mechanisms raises questions about their effectiveness and the limited number of studies that evaluate them in depth, as well as how to establish the choice of techniques in future studies. Table 4 shows the ML models, along with their frequency, references, and model type, arranged in descending order.

In Table 4, we highlight the architecture based on pre-trained transformers, with BERT being the most widely used model (20 studies), followed by SVM, a discriminative classifier, and RoBERTa, another pre-trained transformer (11 studies each). We also observe a high presence of traditional models, such as Random Forest, a tree-based classifier, and BiLSTM, a recurrent neural network (each mentioned 10 times), as well as Naive Bayes, a probabilistic classifier. Likewise, LSTM, another recurrent neural network (9 studies), is designed to process sequential data with temporal dependencies, and CNN, a convolutional neural network (7 studies), is designed to process data with spatial structure. All of the above reflects their validity in tasks of emotional tone classification in hate speech.

 doi:10.20944/preprints202510.2435.v1

Table 4. Machine Learning Models.

Technique	Articles/Authors	Category	Frequency
BERT	[3,4,14,16,18–21,23,24,27–29,31,34–36,39–41]	Pre-trained Transformer	20
RoBERTA	[3,4,19,21,27–29,34–36]	Pre-trained Transformer	11
SVM	[3,5,17,21,22,25,27,30–32,36]	Discriminative Classifier	11
BiLSTM	[3–5,7,15,18,19,26,31,40]	Recurrent Neural Network (RNN)	10
Random Forest	[5,15,17,20,22,27,30–32,36]	Tree-Based Classifier	10
LSTM	[3,17,19,26,30–32,39,40]	Recurrent Neural Network (RNN)	9
Naive Bayes	[5,15,17,19,21,22,27,30,31]	Probabilistic Classifier	9
CNN	[5,17,19,26,31,39,40]	Convolutional Neural Network	7
GPT-3	[6,33,41]	Large Language Model (LLM)	3
GPT-4	[6,33,41]	Large Language Model (LLM)	3
Gemini	[6,33,41]	Large Language Model (LLM)	3
LLaMA 2	[33,34,41]	Large Language Model (LLM)	3
Gemma	[23,33]	Large Language Model (LLM)	2
Total			**101**

It is worth highlighting more recent models, such as GPT-3, GPT-4, LLaMA, LLaMA 2, Gemma, and Gemini, all of which are generative models for training large amounts of text (LLMs), which appear less frequently (in 2 or 3 studies). However, they are showing emerging use, especially in zero-shot or few-shot approaches, which highlights their potential in tasks with less labeled data. These results indicate a growing trend toward pre-trained, contextual language models without neglecting the validity of traditional models due to their simplicity and interpretability. Figure 2 presents a visual representation of the results summary for RQ1, facilitating the identification of the NLP techniques and ML models most used in classifying emotional tones in hate speech [42].

RQ2: *What tools do authors use to implement NLP and machine learning techniques for detecting hate speech and/or emotional tone on social media?*

Given this research question, the uniform implementation of preprocessing with NLTK and Spacy addresses the need to manage the variability and noise inherent in social media language. Rodriguez et al. [25] highlights that NLTK facilitates the tokenization and elimination of stopwords, as well as the integration of custom dictionaries to filter hashtags and URLs. On the other hand, Chakraborty et al. [17] note that Spacy accelerates the process with its architecture, which is based on static vectors that enable efficient lemmatization of large volumes of text, thereby reducing computational costs compared to other libraries.

 doi:10.20944/preprints202510.2435.v1

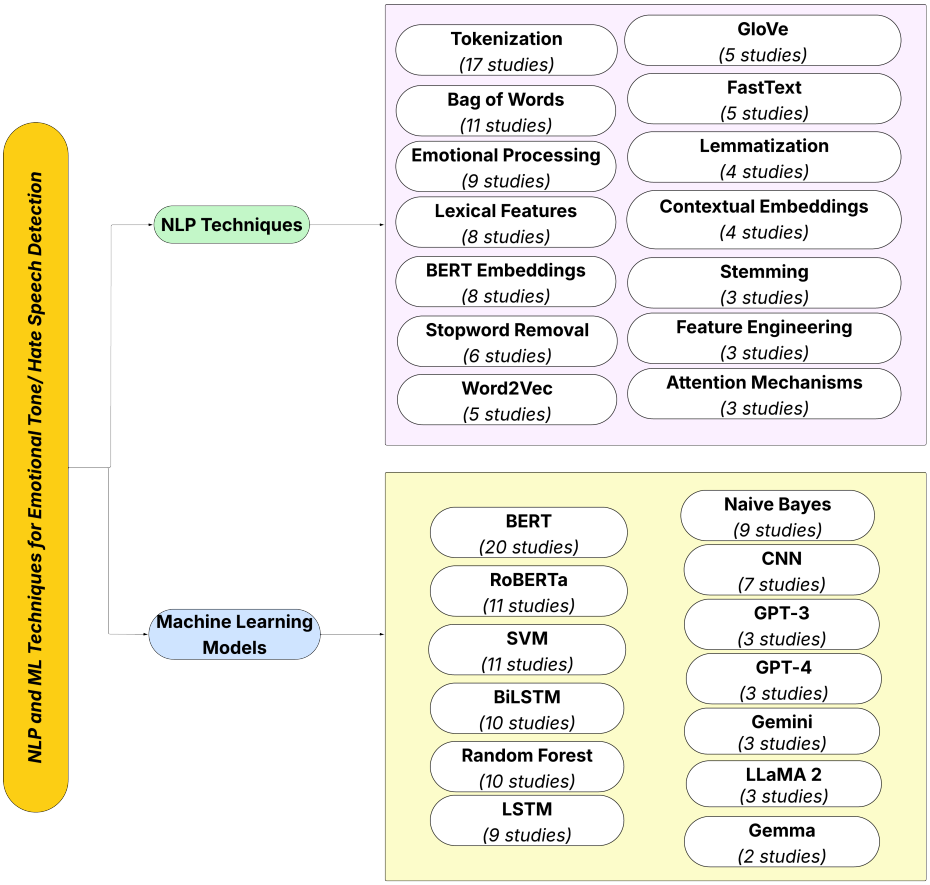

Figure 2. Most frequent techniques and models in the classification of the emotional tone of hate speech using ML and NLP.

For static lexical representation, the authors opted for Word2Vec and FastText via Gensim, not only because of their ease of use but because both models can be quickly trained on specialized corpora. Koufakou et al. [21] showed that Word2Vec effectively captures fine-grained semantic relationships between offensive terms and their contexts. Finally, Alaeddini [30] utilizes FastText to handle mixed code (Hindi-English) cases and neologisms robustly, thereby improving the coverage of rare or emerging vocabulary.

The leap towards contextual embedding is supported by the adoption of Hugging Face's Transformers library. Alvarez-Gonzalez et al. [15] compare BERT and RoBERTa in emotion detection tasks on Twitter and report that these models increase the F1-score by up to 7% compared to Word2Vec. The authors place particular emphasis on disambiguating irony and double entendres (i.e., two interpretations) common in hate speech. Ngo and Kocoń [31] reinforce this finding by combining contextual embeddings with user features to improve multimodal classification.

In the modeling phase, however, the preference for Scikit-learn (SVM, Random Forest, Naive Bayes) is due to its stability and transparency; this preference was valued in studies that require interpretability, such as Min et al. [3] when designing a multi-label self-training scheme where Scikit-learn's cross-validation yields reproducible results.

On the other hand, TensorFlow/Keras and PyTorch are the core of deep architectures in the articles by Al-Hashedi et al. [14], in which they implement a CNN–BiLSTM with attention and emotional lexicon in Keras. Paul et al. [4] developed a multimodal Transformer in PyTorch with cross-attention that integrates text, emojis, and metadata.

Figure 3 shows the frequency of use of each tool, with the upper end dominated by the Natural Language Toolkit (30 studies) and spaCy (28 studies), which serve as pillars of preprocessing, ensuring the cleaning and normalization of the text before any other phase. Subsequently, the bars for Scikit-

 doi:10.20944/preprints202510.2435.v1

learn (22 studies) and Transformers (20 studies) indicate the transition to feature extraction and the comparison between classical methods and advanced embeddings. In the middle positions, Word2Vec (15 studies) and FastText (12 studies) confirm their complementary role as static representations. At the same time, the shorter TensorFlow/Keras (18 studies) and PyTorch (5 studies) show that deep architecture is only applied when more complex modeling is required.

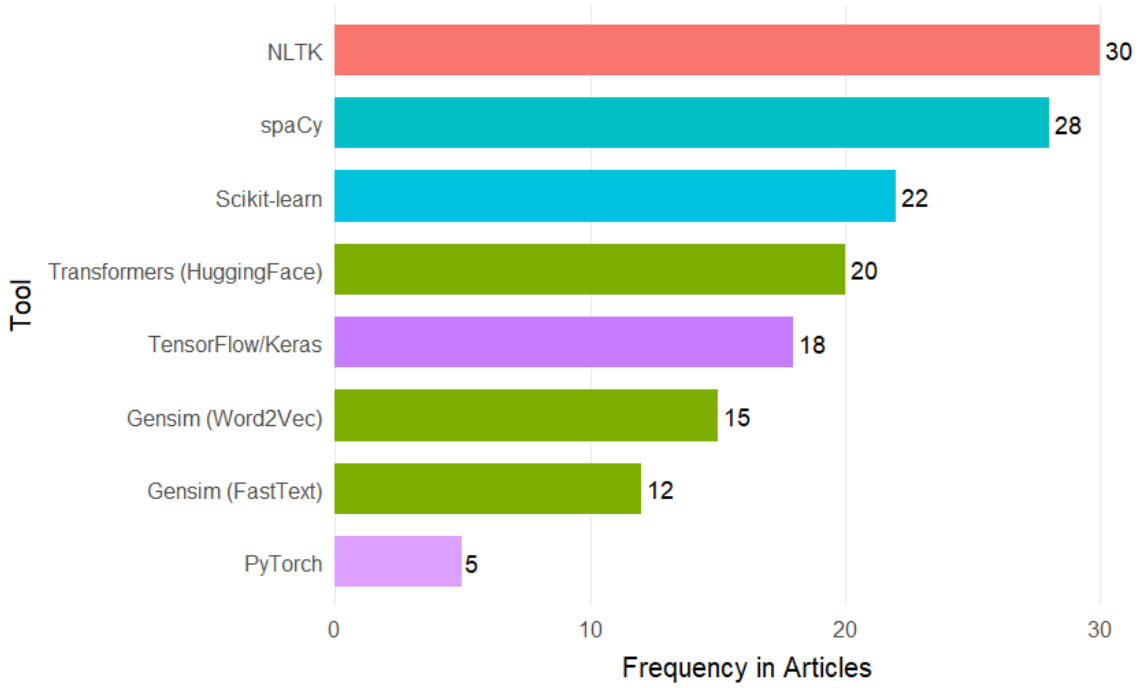

Figure 3. Most frequently used tools to detect emotional tone in hate speech identified in the SLR.

RQ3: *What are the most detected emotions in hate speech?*

The emotions most frequently detected in hate speech are consistent with the study proposed by psychologist Paul Ekman. Ekman developed the theory of basic emotions, which recognizes the universal existence of six emotions: joy, sadness, anger, fear, surprise, and disgust. These emotions are the most used in computer science research [43], which are strongly perceived and easily distinguishable.

In this context, we highlight that the most frequently detected emotions are anger (22 studies), followed by fear and sadness (each with 20 studies), surprise (18 studies), and disgust (17 studies). These represent responses to basic emotions in the face of threats or conflict situations and frequently appear in hate speech. This trend is supported by studies such as that of Kastrati et al. [19]. They labeled 17.5 million tweets using distant supervision with emojis aligned with the Ekman model and found that anger, sadness, disgust, and fear were the most recurrent emotions in hate speech. Similarly, the study by Min et al. [3] analyzed three hate speech datasets and concluded that the same emotions were dominant when classifying messages with violent and discriminatory content.

Likewise, in this study, we identified emotions considered positive, such as happiness (15 studies), love, and gratitude (each in 6 studies). The presence of these emotions may be related to the ironic or ambiguous use of language, which is common in contexts where affective expressions are used to reinforce hostile discourse indirectly. As mentioned by Baruah et al. [5] and Al-Hashedi et al. [14], where speeches are analyzed that, although they seem positive, are loaded with hate or double meanings. Although these emotions do not dominate automatic analysis, they enable us to explore the sarcastic dimensions of discourse.

Finally, more than 40 different emotions were detected, demonstrating the emotional complexity behind hate speech, mainly when directed at vulnerable groups. The presence of negative emotions reinforces the notion that this type of speech is characterized by hostility, threat, and rejection. By

 doi:10.20944/preprints202510.2435.v1

studying these emotions, we find them extremely useful for training emotional classification models. The distribution of emotions found in this study is presented in Figure 4.

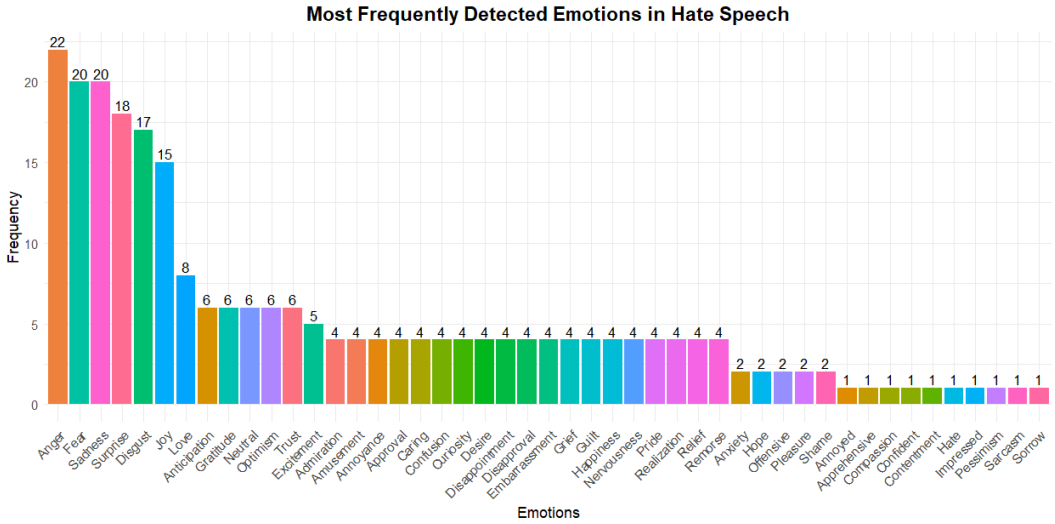

Figure 4. Distribution of the most frequently identified types of emotions.

RQ4: *What are the main challenges, limitations, and future research directions for using NLP and machine learning techniques for detecting emotional tone in hate speech?*

Automating emotional tone detection in hate speech faces a first challenge: emotional polarity. Traditionally, many methods rely on classifying each text as negative, positive, or neutral, but in the context of hate speech, this approach is insufficient. As shown in Figure 5, almost half of the hostile messages (49%) are labeled as negative, but a surprising 37.8% appear with a positive polarity, and 13.3% are labeled as neutral. This distribution indicates the frequency of sarcasm, irony, or "hostility camouflaged" with friendly expressions, causing models based solely on polarity to overlook much of the offensive content [35]. Therefore, a future direction is to incorporate communicative intent annotations and specific irony detection metrics to reduce false negatives and enrich contextual analysis.

Another critical challenge is the scarcity of high-quality, multilingual data. Most work focuses on English-language corpora, which biases and limits the validity of the systems in other languages and dialects [1]. Without balanced repositories that include minority languages and unified annotation schemes for both hate speech categories and emotional nuances, it is tough to compare approaches and ensure cultural equity. As a future solution, it is essential to promote collaborative initiatives that create multilingual repositories with consistent annotations and quality protocols, thereby facilitating reproducibility and benchmarking.

The ambiguity and subtlety of language constitute another barrier. Expressions laden with sarcasm, youth slang, or double entendres (i.e., two interpretations) often fool the most advanced contextual models. Fillies and Paschke [35] demonstrate that even in transformers like BERT, detecting irony and emerging neologisms yields high error rates. To mitigate this, evaluation frameworks are needed that include "youth bias" or "sarcasm detection" metrics, as well as communicative intent annotations that allow algorithms to differentiate between message form and function.

On the technical side, the adoption of deep learning models is hampered by their high computational demands and opacity. Architectures such as those based on multimodal Transformers require substantial volumes of labeled data and computing power, making them challenging to utilize in resource-constrained environments Wang et al. [44]. Furthermore, their "black box" nature prevents auditing and explaining decisions, a critical aspect in sensitive applications. Future directions include model distillation into lightweight versions and the integration of explainable AI (XAI) techniques that justify predictions and facilitate adoption in contexts of high social responsibility.

 doi:10.20944/preprints202510.2435.v1

Finally, a multimodal fusion of text, images, emojis, and user metadata will enrich the emotional analysis, as demonstrated by Paul et al. [4], who incorporate cross-attention in Twitter. However, these approaches require synchronized and labeled datasets across modalities, which are still scarce in the literature. Future work should focus on developing multimodal annotation protocols, exploring temporal alignment methods, and studying how to efficiently combine signals from different channels to capture emotional tones more holistically.

Figure 5 presents a visualization that empirically reinforces these findings. The results indicate that the predominant emotional polarity in hate speech is negative (49%), followed by a surprising proportion of positively polarized messages (37.8%) and a minority of neutral messages (13.3%).

The pattern in Figure 5 is consistent with what is expected, given that hate speech is often charged with emotions such as contempt, anger, or rejection. However, the high proportion of messages classified as positive or neutral points to another dimension of the problem. Many hostile messages are crafted to appear benign or employ emotionally positive grammatical structures that conceal harmful content. This is consistent with the argument by Fillies and Paschke [35], who argue that contemporary language, particularly among young people and on social media, includes forms of expression that are difficult to capture by outdated or biased models. The presence of these disguised emotions underscores the need to incorporate emotional detection as an integral part of automated moderation systems.

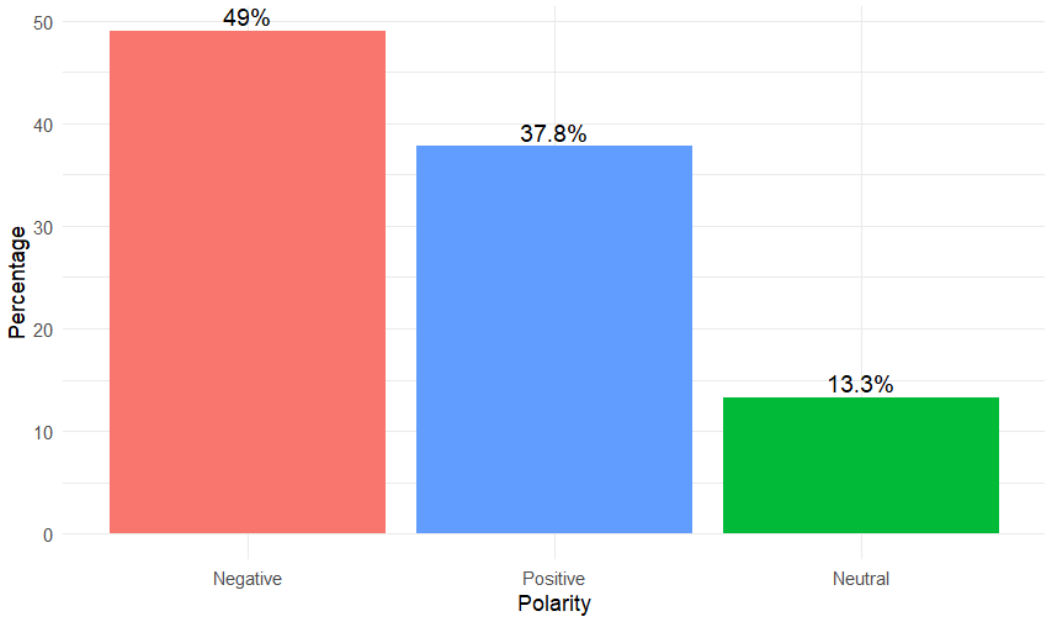

Figure 5. Percentage of Polarity of Emotions in Hate Speech.

Furthermore, the possibility that hate speech may be perceived as emotionally neutral also raises methodological questions.Ramos et al. [1] suggest that emotion should be assessed in conjunction with the full communicative context rather than solely from the plain text. Consequently, the detection of emotional tone in hate speech should consider not only polarity but also its implicit manifestation, its temporal evolution, and its relationship to the conversational environment. In this sense, the contributions of Wang et al. [44] on multimodal models open the door to a more holistic and nuanced detection, allowing us to understand not only the content but also the intention and emotional impact of messages.

RQ5: *Which NLP or machine learning models perform best in emotional tone classification based on metrics such as precision, recall, or F1 score?*

In analyzing the primary studies, researchers used the F1 score, a metric that represents the harmonic meaning between precision and recall, to detect emotional tone in hate speech. This measure

 doi:10.20944/preprints202510.2435.v1

provides a balanced evaluation when working with datasets that have a class imbalance. A value close to 1 (or 100%) indicates that the model correctly identifies and retrieves relevant instances.

Based on the F1-score analysis, the LLaMA 2 model achieved 100% performance on this metric, positioning itself as the most practical for classifying emotional tone in hate speech. This result was obtained by applying context-based learning to five examples within the Google Play dataset, which features a small number of labeled records and an uneven distribution between classes.

The LLaMA 2 model operates on a transformer-like architecture that implements RMSNorm normalization, employs the SwiGLU activation function, and utilizes rotational position encoding. It was trained with trillions of tokens and can handle long inputs, which facilitates the interpretation of broad contexts. The model does not require weight tuning, as it generates predictions only from examples within the same input message; this enables its application in scenarios where there is insufficient data for supervised training.

On the other hand, HingRoBERTa, which specializes in Hindi-English code-mixed texts, follows with a 98.45% score [28]. LSTM [17] and BERT combined with EDM/NRC (Emotion Detection Model and NRC Emotion Lexicon) [14], both with a 97% score, reflecting the usefulness of classical models and lexical techniques in specific contexts.

Likewise, other models achieved outstanding results, such as BiLSTM with attention mechanisms, which scored 96.45% Li et al. [7], DistilRoBERTa, which scored 94.4% [31], and G-BERT, which scored 92.15% [20]. These values show a preference for Transformer-based models and combinations with emotion-focused attention techniques or embeddings, which demonstrate the ability to capture emotional context in language. Figure 6 presents the best-performing models based on their average F1 score, highlighting the effectiveness of modern architecture compared to classical approaches.

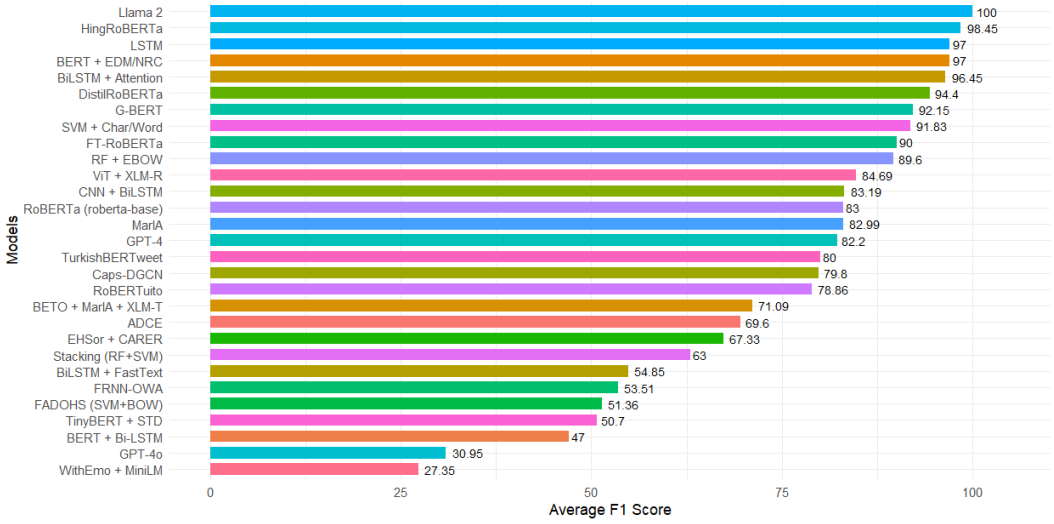

Figure 6. Models identified with the best results with the F1-Score metric for the detection of emotional tone in hate speech.

4. Discussion

This study presents an analysis of the existing literature on the classification of emotional tone in hate speech, utilizing natural language processing and machine learning techniques. Every aspect was considered, from the methodology to the results obtained. Through a selective search and review of publications available on IEEE Xplore, ScienceDirect, ACM, Springer, and Wiley digital databases, we identified 82 articles initially. Applying the inclusion and exclusion criteria, we selected 34 relevant primary studies from these sources.

Firstly, we examine the most frequently used natural language processing techniques. These include tokenization, the most common approach, bag-of-words, emotional processing, BERT, Word2vec, GloVe, FastText, lemmatization, stop-word removal, and attention mechanisms. These techniques

 doi:10.20944/preprints202510.2435.v1

facilitate the preparation and transformation of texts to identify patterns that aid in detecting emotional tone in hate speech. We see a growing trend toward approaches that combine traditional processing with context-based deep learning models. This methodological evolution demonstrates that the field is migrating toward more complex models that not only analyze the surface of the text but also incorporate deeper semantic elements.

Regarding machine learning models, we identified a higher frequency in Transform-based architecture, with BERT being the most used, followed by its variant RoBERTa. We also identified traditional models, such as SVM, Random Forest, and Naive Bayes, as well as neural network models, including CNN and BiLSTM. Additionally, we explored the use of models like GPT-3, GPT-4, LLaMA 2, Gemini, and Gemma. Although these were less frequently used, they reflect the potential in tasks where labeled data is scarce, and they can apply learned knowledge in other contexts (i.e., zero-shot or few-shot learning). Recently, advanced and pre-trained models that have been adopted have shown promising results. However, it is essential to investigate further why some less-used yet high-performing models, such as LLaMA 2, have not yet been widely adopted. This could be due to factors such as code availability, implementation difficulty, or lack of external validation.

It is important to note that some of the best-performing models are not the most commonly used in reviewed studies, demonstrating that frequency of use is not always directly related to their effectiveness. Among the models with the highest F1 scores are LlaMA 2 (100%), HingRoBERTa (98.45%), and models such as LSTM and BERT+EDM/NRC (97% each). These results indicate that less-cited models can outperform the most popular ones. This finding highlights the need for future research to be more inclusive of replicating well-known architecture; instead, recent proposals that have demonstrated greater generalization and accuracy should be explored.

Regarding the identified emotions, we find an alignment with Ekman's theory of emotions. When comparing negative and positive emotions, the former predominates. Among them, anger, fear, and sadness stand out as the most frequent, which is consistent with the hostile nature of this type of discourse. When discussing positive emotions, we encounter happiness, love, and gratitude, which, although ostensibly positive and non-harmful, can be associated with the use of ironic or ambiguous expressions, adding complexity to the analysis and highlighting the need to improve approaches that can capture less explicit semantic nuances. Along these lines, Fillies and Paschke [35] highlight those hostile expressions disguised as positive are increasingly frequent in youth speech, representing a considerable challenge for models that work solely with emotional polarity, thus generating a high number of false negatives.

On the other hand, we underline a series of challenges and limitations in the use of NLP and ML techniques for detecting emotional tone in hate speech. One of the primary issues is linguistic and cultural bias, as evidenced by the predominance of models trained in English or Arabic. This highlights a lack of multilingual databases, which limits the models' generalization capabilities. Furthermore, the ambiguity of language in expressions such as irony or sarcasm, which are increasingly common among young people, represents an obstacle, as they often elude traditional detection mechanisms. This situation highlights the urgent need to build multilingual and culturally diverse corpora with more refined emotional annotations, allowing for the training of models that are adaptable to different sociolinguistic contexts.

Among the limitations inherent to machine learning models for detecting emotional tone in hate speech using ML and NLP, we identify those based on deep learning, such as BERT, RoBERTa, GPT-3/4, BiLSTM, and CNN, which are computationally intensive, making them difficult to implement in low-resource environments. Furthermore, since they are "black box" models, the interpretability of the results is reduced, which is a fundamental aspect in the detection of hate speech, where transparency is required to avoid unfair or arbitrary decisions. Some models can repeat biases that already exist in the data and, therefore, end up treating certain groups poorly or unfairly. In this sense, it becomes essential to incorporate Explainable Artificial Intelligence (XAI) techniques that allow the models' decisions to be audited and ensure their behavior is ethical and responsible.

 doi:10.20944/preprints202510.2435.v1

Similarly, despite its widespread adoption, NLTK has efficiency limitations, being considerably slower than modern tools like spaCy in basic preprocessing tasks. However, spaCy sacrifices flexibility in favor of performance, which limits its adaptation to experimental contexts. Meanwhile, TensorFlow/Keras, although powerful, can be complex to debug and excessive in scenarios where deep modeling is not required, which limits its use in studies with limited computational resources.

Finally, we emphasize that the findings of this systematic review offer an insight into the current state of the techniques and models most commonly used for detecting emotional tone in hate speech, which highlights the need to continue improving tools to address the complexity of language in the digital environment to make hate speech safer and more respectful. In this regard, it is recommended that future research advance toward the development of hybrid models that combine supervised approaches with deep semantic interpretation capabilities. Additionally, we suggested cross-validating models in different regions, designing multimodal annotations that incorporate text, emojis, and user context, and integrating systems that not only detect emotions but also interpret the communicative intent behind messages.

5. Conclusions and Future Work

This study applied a combination of PRISMA and PICOS methodological guides to conduct a systematic literature review, highlighting the importance of natural language processing and machine learning in detecting emotional tones in hate speech. The findings show that among the 34 primary studies analyzed, these techniques enable the identification of emotional patterns targeting vulnerable groups. Among the most widely used are those focused on linguistic preprocessing and semantic embeddings, such as Word2Vec and BERT. We observed a preference for models such as BERT, RoBERTa, SVM, and Random Forest, although some of the fewer common ones, such as LLaMA 2, demonstrated superior performance. We also identified a clear predominance of negative emotions such as anger and fear, consistent with the hostile nature of this type of speech, as well as the ironic use of positive emotions, which poses significant semantic challenges. The main limitations of our study include the scarcity of multilingual corpora and the difficulties in interpreting ambiguous expressions such as sarcasm.

As future work, we plan to develop hybrid models that integrate supervised techniques with deep semantic analysis, aiming for more precise, contextualized, and respectful emotional detection.

Author Contributions: Conceptualization, Aymé Escobar Díaz, Ricardo Rivadeneira, and Walter Fuertes; Methodology, Aymé Escobar Díaz and Ricardo Rivadeneira; Literature search and selection, Aymé Escobar Díaz and Ricardo Rivadeneira; Validation, Aymé Escobar Díaz, Ricardo Rivadeneira and Walter Fuertes; Formal analysis, Aymé Escobar Díaz, Ricardo Rivadeneira and Walter Fuertes; Investigation, Aymé Escobar Díaz and Ricardo Rivadeneira; Data curation, Aymé Escobar Díaz and Ricardo Rivadeneira; Writing—original draft preparation, Aymé Escobar Díaz and Ricardo Rivadeneira; Writing—review and editing, Walter Fuertes; Visualization, Aymé Escobar Díaz and Ricardo Rivadeneira; Supervision, Walter Fuertes; Project administration, Walter Fuertes; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The Universidad de las Fuerzas Armadas ESPE will fund the Article Processing Charge (APC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created in this study. The data supporting the findings of this systematic literature review consist of previously published articles that are publicly available and cited throughout the manuscript. Therefore, no additional data are available.

Acknowledgments: The authors express their sincere gratitude to the Universidad de las Fuerzas Armadas ESPE for the academic, technical, and institutional support provided for the development of this research. During the preparation of this manuscript, OpenAI ChatGPT (GPT-5, 2025) was used to assist with style correction and LaTeX

 doi:10.20944/preprints202510.2435.v1

formatting. The authors have reviewed and edited all generated content and take full responsibility for the final version of the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Ramos, G.; Batista, F.; Ribeiro, R.; Fialho, P.; Moro, S.; Fonseca, A.; Guerra, R.; Carvalho, P.; Marques, C.; Silva, C. A comprehensive review on automatic hate speech detection in the age of the transformer. *14*. Publisher: Springer Science and Business Media LLC, https://doi.org/10.1007/s13278-024-01361-3.
2. Zhang, F.; Chen, J.; Tang, Q.; Tian, Y. Evaluation of emotion classification schemes in social media text: an annotation-based approach. *12*, 503. Publisher: Springer Nature, https://doi.org/10.1186/s40359-024-01665-1.
3. Min, C.; Lin, H.; Li, X.; Zhao, H.; Lu, J.; Yang, L.; Xu, B. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *96*, 214–223. Publisher: Elsevier BV, https://doi.org/10.1016/j.inffus.2023.03.015.
4. Paul, J.; Mallick, S.; Mitra, A.; Roy, A.; Sil, J. Multi-modal Twitter Data Analysis for Identifying Offensive Posts Using a Deep Cross-Attention–based Transformer Framework. *19*, 1–30. Publisher: Association for Computing Machinery (ACM), https://doi.org/10.1145/3713077.
5. Baruah, A.; Wahlang, L.; Jyrwa, F.; Shadap, F.; Barbhuiya, F.; Dey, K. Abusive Language Detection in Khasi Social Media Comments. Publisher: Association for Computing Machinery (ACM), https://doi.org/10.1145/3664285.
6. Kastrati, M.; Imran, A.S.; Hashmi, E.; Kastrati, Z.; Daudpota, S.M.; Biba, M. Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with Explainable Artificial Intelligence. *149*, 110136. Publisher: Elsevier BV, https://doi.org/10.1016/j.engappai.2025.110136.
7. Li, Y.; Chan, J.; Peko, G.; Sundaram, D. An explanation framework and method for AI-based text emotion analysis and visualisation. *178*, 114121. Publisher: Elsevier BV, https://doi.org/10.1016/j.dss.2023.114121.
8. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *74*, 790–799. Publisher: Elsevier BV, https://doi.org/10.1016/j.recesp.2021.06.016.
9. Frandsen, T.F.; Bruun Nielsen, M.F.; Lindhardt, C.L.; Eriksen, M.B. Using the full PICO model as a search tool for systematic reviews resulted in lower recall for some PICO elements. *127*, 69–75. Publisher: Elsevier BV, https://doi.org/10.1016/j.jclinepi.2020.07.005.
10. Zapata, J.I.; Garcés, E.; Fuertes, W. Ransomware Detection with Machine Learning: Techniques, Challenges, and Future Directions - A Systematic Review. *15*, 271–287. Publisher: SASA Publications, https://doi.org/10.58346/jisis.2025.i1.017.
11. Macas, M.; Wu, C.; Fuertes, W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Systems with Applications* **2024**, *238*, 122223. https://doi.org/10.1016/j.eswa.2023.122223.
12. Benavides-Astudillo, E.; Fuertes, W.; Sanchez-Gordon, S.; Nuñez-Agurto, D.; Rodríguez-Galán, G. A Phishing-Attack-Detection Model Using Natural Language Processing and Deep Learning. *Applied Sciences* **2023**, *13*, 5275. https://doi.org/10.3390/app13095275.
13. Haynes, R.B. Forming research questions. *59*, 881–886. Publisher: Elsevier.
14. Al-Hashedi, M.; Soon, L.K.; Goh, H.N.; Lim, A.H.L.; Siew, E.G. Cyberbullying Detection Based on Emotion. *11*, 53907–53918. Publisher: Institute of Electrical and Electronics Engineers (IEEE), https://doi.org/10.1109/access.2023.3280556.
15. Alvarez-Gonzalez, N.; Kaltenbrunner, A.; Gómez, V. Uncovering the Limits of Text-based Emotion Detection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.219.
16. Bashynska, I.; Sarafanov, M.; Manikaeva, O. Research and Development of a Modern Deep Learning Model for Emotional Analysis Management of Text Data. *14*, 1952. Publisher: MDPI AG, https://doi.org/10.3390/app14051952.
17. Chakraborty, P.; Nawar, F.; Chowdhury, H.A. Sentiment Analysis of Bengali Facebook Data Using Classical and Deep Learning Approaches. In *Innovation in Electrical Power Engineering, Communication, and Computing Technology*; Springer Singapore; pp. 209–218. ISSN: 1876-1119, https://doi.org/10.1007/978-981-16-7076-3_19.

 doi:10.20944/preprints202510.2435.v1

18. de León Languré, A.; Zareei, M. Improving Text Emotion Detection Through Comprehensive Dataset Quality Analysis. *12*, 166512–166536. Publisher: Institute of Electrical and Electronics Engineers (IEEE), https://doi.org/10.1109/access.2024.3491856.
19. Kastrati, M.; Kastrati, Z.; Shariq Imran, A.; Biba, M. Leveraging distant supervision and deep learning for twitter sentiment and emotion classification. *62*, 1045–1070. Publisher: Springer Science and Business Media LLC, https://doi.org/10.1007/s10844-024-00845-0.
20. Keya, A.J.; Kabir, M.M.; Shammey, N.J.; Mridha, M.F.; Islam, M.R.; Watanobe, Y. G-BERT: An Efficient Method for Identifying Hate Speech in Bengali Texts on Social Media. *11*, 79697–79709. Publisher: Institute of Electrical and Electronics Engineers (IEEE), https://doi.org/10.1109/access.2023.3299021.
21. Koufakou, A.; Garciga, J.; Paul, A.; Morelli, J.; Frank, C. Automatically Classifying Emotions based on Text: A Comparative Exploration of Different Datasets. In Proceedings of the 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 342–346. https://doi.org/10.1109/ictai56018.2022.00056.
22. Liapis, C.M.; Karanikola, A.; Kotsiantis, S. Enhancing sentiment analysis with distributional emotion embeddings. *634*, 129822. Publisher: Elsevier BV, https://doi.org/10.1016/j.neucom.2025.129822.
23. Pan, R.; García-Díaz, J.A.; Valencia-García, R. Spanish MTLHateCorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity. *94*, 103990. Publisher: Elsevier BV, https://doi.org/10.1016/j.csi.2025.103990.
24. Priya, P.; Firdaus, M.; Ekbal, A. A multi-task learning framework for politeness and emotion detection in dialogues for mental health counselling and legal aid. *224*, 120025. Publisher: Elsevier BV, https://doi.org/10.1016/j.eswa.2023.120025.
25. Rodriguez, A.; Chen, Y.L.; Argueta, C. FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis. *10*, 22400–22419. Publisher: Institute of Electrical and Electronics Engineers (IEEE), https://doi.org/10.1109/access.2022.3151098.
26. Sasidhar, T.T.; B, P.; P, S.K. Emotion Detection in Hinglish(Hindi+English) Code-Mixed Social Media Text. *171*, 1346–1352. Publisher: Elsevier BV, https://doi.org/10.1016/j.procs.2020.04.144.
27. Sohail, T.; Aiman, A.; Hashmi, E.; Imran, A.S.; Daudpota, S.M.; Yayilgan, S.Y. Hate Speech Detection in Code-Mixed Datasets Using Pretrained Embeddings and Transformers. In Proceedings of the 2024 International Conference on Frontiers of Information Technology (FIT). IEEE, pp. 1–6. https://doi.org/10.1109/fit63703.2024.10838452.
28. Takawane, G.; Phaltankar, A.; Patwardhan, V.; Patil, A.; Joshi, R.; Takalikar, M.S. Language augmentation approach for code-mixed text classification. *5*, 100042. Publisher: Elsevier BV, https://doi.org/10.1016/j.nlp.2023.100042.
29. Vallecillo-Rodríguez, M.E.; Plaza-del Arco, F.M.; Montejo-Ráez, A. Combining profile features for offensiveness detection on Spanish social media. *272*, 126705. Publisher: Elsevier BV, https://doi.org/10.1016/j.eswa.2025.126705.
30. Alaeddini, M. Emotion Detection in Reddit: Comparative Study of Machine Learning and Deep Learning Techniques.
31. Ngo, A.; Kocoń, J. Integrating personalized and contextual information in fine-grained emotion recognition in text: A multi-source fusion approach with explainability. *118*, 102966. Publisher: Elsevier BV, https://doi.org/10.1016/j.inffus.2025.102966.
32. Touahri, I.; Mazroui, A. Enhancement of a multi-dialectal sentiment analysis system by the detection of the implied sarcastic features. *227*, 107232. Publisher: Elsevier BV, https://doi.org/10.1016/j.knosys.2021.107232.
33. Lecourt, F.; Croitoru, M.; Todorov, K. " Only ChatGPT gets me": An Empirical Analysis of GPT versus other Large Language Models for Emotion Detection in Text.
34. Zhang, T.; Irsan, I.C.; Thung, F.; Lo, D. Revisiting Sentiment Analysis for Software Engineering in the Era of Large Language Models. *34*, 1–30. Publisher: Association for Computing Machinery (ACM), https://doi.org/10.1145/3697009.
35. Fillies, J.; Paschke, A. Youth language and emerging slurs: tackling bias in BERT-based hate speech detection. Publisher: Springer Science and Business Media LLC, https://doi.org/10.1007/s43681-025-00701-z.
36. García-Díaz, J.A.; Cánovas-García, M.; Colomo-Palacios, R.; Valencia-García, R. Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *114*, 506–518. Publisher: Elsevier BV, https://doi.org/10.1016/j.future.2020.08.032.

 doi:10.20944/preprints202510.2435.v1

37. Kaminska, O.; Cornelis, C.; Hoste, V. Fuzzy rough nearest neighbour methods for detecting emotions, hate speech and irony. *625*, 521–535. Publisher: Elsevier BV, https://doi.org/10.1016/j.ins.2023.01.054.
38. Ankita.; Rani, S.; Bashir, A.K.; Alhudhaif, A.; Koundal, D.; Gunduz, E.S. An efficient CNN-LSTM model for sentiment detection in #BlackLivesMatter. *193*, 116256. Publisher: Elsevier BV, https://doi.org/10.1016/j.eswa.2021.116256.
39. Mittal, U. Detecting Hate Speech Utilizing Deep Convolutional Network and Transformer Models. In Proceedings of the 2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM). IEEE, pp. 1–4. https://doi.org/10.1109/elexcom58812.2023.10370502.
40. Zhu, X.; Lou, Y.; Deng, H.; Ji, D. Leveraging bilingual-view parallel translation for code-switched emotion detection with adversarial dual-channel encoder. *235*, 107436. Publisher: Elsevier BV, https://doi.org/10.1016/j.knosys.2021.107436.
41. Najafi, A.; Varol, O. TurkishBERTweet: Fast and reliable large language model for social media analysis. *255*, 124737. Publisher: Elsevier BV, https://doi.org/10.1016/j.eswa.2024.124737.
42. Andrade, R.O.; Fuertes, W.; Cazares, M.; Ortiz-Garcés, I.; Navas, G. An Exploratory Study of Cognitive Sciences Applied to Cybersecurity. *Electronics* **2022**, *11*, 1692. https://doi.org/10.3390/electronics11111692.
43. Yadollahi, A.; Shahraki, A.G.; Zaiane, O.R. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *50*, 1–33. https://doi.org/10.1145/3057270.
44. Wang, S.; Shibghatullah, A.S.; Iqbal, T.J.; Keoy, K.H. A review of multimodal-based emotion recognition techniques for cyberbullying detection in online social media platforms. *36*, 21923–21956. Publisher: Springer, https://doi.org/10.1007/s00521-023-08766-5.