National Research Journal of Information Technology & Information Science
Volume No: 13, Issue No: 2, Year: 2026 (July-December)
PP: 86-102
ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

TRANSFORMER-BASED MODELS FOR HATE SPEECH DETECTION: A SYSTEMATIC REVIEW

Inderpreet Kaur

Research Scholar, School of Engineering & Technology, CT University, Ludhiana, Punjab, India

Rupali Dhir

Associate Professor, Department of CSE, CT University, Ludhiana-142024, Punjab, India

ABSTRACT

Hate speech on social media has grown fast enough that manual moderation alone can no longer keep up with it, which has pushed researchers toward automatic detection systems built on transformer models. This review studies the evolution of hate speech detection architectures starting from general-purpose encoders like BERT and RoBERTa to domain-specific variants such as HateBERT and BETO to large language models like GPT-3/4 and LLaMA. Various databases such as ACL Anthology, IEEE Xplore, Springer and arXiv were searched and relevant papers were identified. The papers were included if they reported empirical results for a transformer or LLM-based classifier on a hate speech or offensive language dataset.The surveyed studies show that domain-specific encoders and ensembles of transformer models generally beat single general-purpose models. When sufficient labelled data is provided, fine-tuned encoders still tend to outperform prompted LLMs, though this gap narrows or reverses in low-resource and code-mixed settings.Persistent problems in hate speech detection include class imbalance, limited explainability and heavy skew toward English datasets. In conclusion, future advancements in this domain include fine-tuning open LLMs on augmented low-resource data and systematically integrating explainability as a foundational architecture rather than a secondary consideration.

Keywords: Hate speech detection, transformer models, BERT, large language models, deep learning, natural language processing, multilingual NLP, text classification.

1. INTRODUCTION

The Internet and social media have really changed how people talk to each other and share things. Using these platforms users freely express their thoughts, emotion and opinions without being limited by distance or traditional restrictions. While this openness helps people share their thoughts freely and promotes global connectivity, it has also given rise to significant problems - one of them being the spread of hate speech.

Hate speech on social media typically targets individuals or groups based on features like gender, caste, religion, politics, race, disability, or sexual orientation. As it is multifaceted and can be expressed in diverse ways and settings, detecting and addressing hate speech remains a highly complex and challenging task. Moreover,the anonymity offered by the Internet can make people feel more free to post harmful and offensive material against specific religions, communities, genders, races and other vulnerable groups. This type of content can threaten social stability, have a negative impact on mental health and on public safety and democratic discourse.

Hate speech in social media has recently become more common. An Ipsos report for Statista, which surveyed the internet in August and September 2023, found that nearly two-thirds of

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

all those who use the internet report that they frequently come across hate speech online. Eight out of ten European Union citizens report they have encountered hate speech on social networks and 40% have personally been threatened or attacked. (Gagliardone et al.,2015) These numbers show how common the problem has become.

During conflicts or crises, it is seen that there is always a sudden spike in hate speech. This could be seen in 2019 when the incident of Christchurch mosque shootings happened in New Zealand which were broadcast live on social media platforms, there was a spike of online racist hate speech. Similarly in 2025, the communal conflicts in Pahalgam, Jammu & Kashmir, resulted in a flood of religiously charged and hateful posts. These events highlight how the interconnection between real-world violence and unrest can fuel online hate speech.

Therefore, there is a need to develop smart and automatic systems to detect and control hate speech in real time.Artificial Intelligence (AI) and Machine Learning (ML) have shown great potential in this area.They enable the analysis of huge amounts of online data and they can be used to effectively create models that will be able to distinguish between hateful and non-hateful speech. As dependency on AI grows,it is essential to ensure accuracy, scalability and usability of these systems.

Different techniques are used to tackle the challenge of detecting hate speech. These are manual features such as the TF-IDF or Bag-of-Words, which are the features commonly used by traditional machine learning approaches like Support Vector Machines (SVM), Naive Bayes and Logistic Regression. They are useful, but do not work when addressing hidden or indirect hate speech because they are not able to understand the actual meaning or context of a sentence.

A traditional machine learning system based on known features like bag-of-words, TF-IDF and n-grams, has been replaced by neural architectures that better model sequence and context and more recently by transformer-based models which have become the state-of-the-art for recent benchmarks. In early studies, classifiers like SVM, Naive Bayes, decision trees and random forests were used, but they failed to capture the contextual and semantic complexity of hateful language (Alrehili, 2019; Schmidt & Wiegand, 2017; Poletto et al., 2021). Recently, deep learning models have emerged in this area. Models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks and transformer-based models like BERT have yielded remarkable results. Next, Vaswani et al(2017) created the Transformer model to learn language models and translation systems that utilizes an attention mechanism where each word can see all the other words, rather than processing the words one to one using an RNN or CNN. This has the advantage of speeding up the training and allowing the model to learn long-range relationships in text in a better way.

When fed with textual inputs, deep learning models learn rich representations, which results in better performance and comparison studies prove that these are generally superior to traditional baselines (Shawkat et al., 2024; Mittal, 2023). Multiple transformer-based models like BERT, RoBERTa, mBERT, XLM-R, MuRIL and HateBERT are now reported to yield state-of-the-art performance, especially in the case of multilingual, subtle and code-mixed hate speech detection, while presenting the challenges of bias, sarcasm, label noise and computational cost (Ramos et al., 2024; Chapagain et al., 2025; Ghosh & Senapati, 2022; Tita & Zubiaga, 2021). These models are able to grasp the meaning and context of the text, which is crucial for dealing with the complexity and volatility of hate speech in online content. This change reflects the increasing use of sophisticated deep learning methods for better and more reliable identification of hate speech.

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

This study examines the current landscape of model adaptation by focusing on four connected research questions:

RQ1: Which models are being used,

RQ2: Which datasets support their development and evaluation,

RQ3: Which fine-tuning strategies are most effective and

RQ4: What open challenges still limit progress.

Together, these questions provide a structured view of both the technical choices and the remaining gaps in the field.

The rest of the paper is organised as follows: Section 2 gives the brief background of the topic covering the aspects such as definition of hate speech, Transformer architecture basics and fine tuning concepts. Literature review is presented in Section 3 which focuses on the related work and recent developments in this field. Section 4 gives the comparative analysis outlining the trends and models used so far. In section 5, challenges and open issues are discussed followed by the future directions and concluding remarks in Section 6 and Section 7 respectively.

2. BACKGROUND

Hate speech can be defined as any insulting public discourse directed at specific individuals or groups of people based on characteristics such as race, religion, ethnicity, nationality, sex, disability, sexual orientation or gender identity. Earlier people used to verbally spread hate speech but now they are using social media platforms such as twitter, facebook etc. to spread hate by posting hateful comments or sharing offensive or abusive content online. There is also an increasing normalisation of abusive behaviour online.

There are various definitions of hate speech in various papers. However, Sellars (2016) does not simply list a few words that should never be used or a few banned words lists, but instead, he looks at what has been done before and "brings to light eight traits" in order to recognize hate speech. Brison (2013) provides a more common normative characterization, as speech that "vilifies, intimidates or incites hatred" against an individual or group based on a perceived social identity. Gelber (2021) further restricts this idea: "It is speech that causes harm, to a degree that justifies government regulation." This makes harm and justification prominent. Similarly, Vilar-Lluch (2023) suggests that hate alone is not a crime; threats, degradation of protected groups, potential harm and intent to incite hatred matter. Another paper authored by Vergani et al. (2024), reveals the ongoing challenges with these definitions, namely differences in the identities that are covered, what constitutes "hateful" conduct and how the "hate element" is evaluated.

One of the main challenges in tackling online hate is that there exists different definitions of hate speech in each country. Various differences in hate speech arise in relation to the various legal frameworks, culture and experiences. This work does not have any specific regional focus, however, it is important to note that the state of hate speech in other countries is not the same. Thus, it is crucial to create flexible, robust and socially aware AI systems.

The transformer models were designed to address issues that previous architectures such as recurrent neural network (RNNs) and long short-term memory (LSTMs) couldn't handle. The previous AI systems used a rather slow process of reading words individually. Also, when they get to the end of the sentence, they will often forget what was there in the beginning. However, in 2017, Transformers were invented, introducing another concept of paying attention to all words in the sentence concurrently and not sequentially (Vaswani et al.,

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

2017). This effect is known as self-attention, this allows the model to determine how relevant each of the other words is to the word that is currently being processed, meaning that context can be anywhere in the sentence and influence the meaning of a single word. The original Transformer had an encoder-decoder structure, being able to read in the input and represent it with the encoder and then use the decoder to produce the output sequence from that representation, for example when translating one sequence to another. BERT, however, only uses the encoder half (Devlin et al., 2019). For a classification problem, you don't need to create new sentences and only need to understand and represent what is already present, which makes it a natural fit for an encoder-only design — and this is the case with hate speech detection, where the task is really to determine the meaning or damage of a piece of text, rather than generating new text. This is the place where the concept of fine tuning comes into play, too. First, BERT is pre-trained on vast amounts of general text, acquiring general patterns of language, grammar and context, without being associated with any specific task. Then fine-tune that pre-trained model and re-train it on a smaller subset of the training data created specifically for a single task, modifying the weights of the model so that it becomes specialized in the task. The model is not starting from scratch and fine-tuning on task-specific data, hence improvement in performance. It has already been trained on a broad language and only requires further training to learn the narrow patterns that are present in a specific type of speech, like the words, tone and context used in hateful speech, which is what Mozafari et al. (2019) did when they fine-tuned BERT for hate speech detection on social media content.

3. LITERATURE REVIEW

3.1 General BERT fine-tuning approaches

After BERT, follow-up work mostly went in two directions. Some studies focused on whether BERT had been trained well enough and others on making it cheaper to run. Liu et al. (2019) proposed an architecture called RoBERTa, which is a replication and refinement of BERT pretraining. The authors systematically tune the most important hyperparameters and the size of the training-set and find that BERT is under-trained and that the optimized model can match or outperform later models. They achieved state-of-the-art results on three datasets namely GLUE, RACE and SQuAD. DistilBERT, a smaller general purpose language model, was introduced by Sanh et al. (2019), suggesting that knowledge distillation can be utilized when training BERT and not just for applying compression to specific tasks. Using a triple loss, which combines language modeling, distillation and cosine-distance objectives, they report this enables BERT to be 40% smaller, to retain 97% of the language understanding capabilities and to accelerate inference by 60%. The paper argues that DistilBERT is a viable alternative for limited compute resources and provides an on-device study to substantiate this claim. Rajput et al. (2021) investigated whether it is possible to achieve state-of-the-art performance in the detection of hate speech without fine-tuning the entire BERT model using only static embeddings, that is, without any fine-tuning of the model's layers. The authors relied on the ETHOS dataset and compared the results of using static BERT embeddings with fastText, GloVe and a combination of fastText/GloVe and found that the neural network produced better results using static BERT embeddings than using the other word representations. They found that static BERT embeddings leads to fewer false positives and that specificity is higher compared to fine-tuned BERT. Mozafari et al. (2019) fine-tune BERT as a transfer-learning model to identify hate speech in two Twitter datasets labeled with offensive content, hate, sexism and racism. They claim that BERT is able to understand the annotation and collection bias better than previous works and believe it is more precise and recall.

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3.2 Domain-specific transformer models

Although pre-trained powerful general purpose language models such as BERT are a major breakthrough, an important development in the field is the move to pre-trained models on corpora of abusive language for specific domains. General models are pre-trained on large and varied text, but much of it is formal. It contains too little of the slang, coded language and implicit aggression that toxic online content relies on. Their lack of exposure restricts their ability to develop a sophisticated understanding of hateful language. To address this deficiency, domain-specific pre-training has been introduced as a way to continue training a model on a corpus consisting of 100% content relevant for the target task, so as to narrow the scope of its vocabulary and contextual understanding. Caselli et al. propose HateBERT, a BERT model that has been retrained with a massive Reddit dataset of banned offensive, abusive and hateful communities. The retrained model outperforms general BERT on three English abusive language datasets and the authors further test its cross-dataset portability, which is dependable on the similarity of the phenomena annotated in the datasets. It can be even more specific, like what led to the development of COVID-HateBERT. The results revealed that models that were trained on non-COVID hate speech corpora were less effective in detecting hate speech about the pandemic. The traditional hateful dataset trained models are not well fitted to COVID-19 related dataset (Li et al., 2021). This required the creation of a new model specialized for a massive collection of COVID-related tweets, demonstrating that topical specialization can be effective even within the context of hate speech.

In addition, a major limitation of general purpose models is that they are not good at detection of non-explicit forms of abuse. Messages containing linguistically subtle and implicit forms of hate speech such as circumlocution, metaphors and sarcasm constitute a real challenge for automatic systems (Ocampo et al., 2023). The complex linguistic constructions are rarely encountered in the formal text used for training the general models. Just like hate speech, the surrounding conversation is also crucial and a weakness of existing strategies for automatic hate speech detection is that it is lacking context (Pérez et al., 2023).

3.3 **Ensemble transformer approaches**

A recent approach to hate speech detection is to build ensembles, which combine transformer models like BERT with classical machine learning classifiers or other neural networks. Each model is trained or fine-tuned on different but related datasets, so their strengths complement each other and the combined prediction is more accurate than any single model's. This also helps with ambiguous or heterogeneous data and it gives more reliable predictions across domains and languages. Common ensemble methods include stacking, soft voting, hard voting and hybrid classifiers that add graph or attention mechanisms.

In the study by Mnassri et al., (2022) they combined publicly available Twitter datasets on the topic of hate speech and offensive speech detection and used ensemble learning methods such as soft voting, max value voting, hard voting and stacking with multiple deep learning models from BERT. The stacking approach has yielded an F1 score as high as 97% on the Davidson dataset and it was found to be effective on the combined dataset as well. This shows the effectiveness of ensemble methods in a variety of imbalanced and diverse hate speech data scenarios. Siddiqui et al., (2024) used multilingual transformers like XLM-RoBerta in combination with explainable AI techniques to tackle issues in multilingual and fine-grained hate speech classification. These ensemble models performed best with high F-scores (up to 91% for binary classification and 86% for fine-grained classes) in all three languages (English, Urdu and Sindhi), clearly demonstrating that ensembling, in complex

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

domainspecific language, is beneficial. Yoo et al.(2025) propose an ensemble of multiple BERT-based models that are trained in parallel for English and Korean hate-speech detection. They found that the ensemble model, based on maximizing the number of votes or prior studies based on probabilistic averaging with a meta-learner, achieves better accuracy than single-model baselines (85% in English and 89% in Korean).

3.4 **Multilingual transformer approaches**

Multilingual transformer models have become a key tool for detecting hate speech in languages that have little labeled data, including those that differ greatly in structure from English. They are based on the transformer architecture models like mBERT, XLM-RoBERTa, MuRIL and IndicBERT which are large-scale pre-trained models trained on multiple languages.

The issue of scarce labeled data for many languages is one of the main challenges of multilingual hate speech detection. To overcome this, a proposal has been made to incorporate a meta learning strategy into models to adapt to new languages when only a few labeled examples are available. For example, the meta learning framework Proto-MAML achieved successful cross-lingual few-shot hate speech and offensive language detection on eight languages and 15 datasets without any manual labor. It outperforms traditional transfer learning approaches by being able to quickly learn to generalize to unseen language even with very few samples (Mozafari et al., 2022). The use of techniques like Bayesian attention with Monte Carlo dropout also helps to improve reliability, because they can offer uncertainty estimates crucial for applications in content moderation that demand trustworthy predictions (Miok et al., 2022). In addition, to support information detection in various languages, hybrid architectures like transformer-CNN or transformer-GRU are proposed to capture both contextual and sequential features for optimizing the detection accuracy (Madhavi et al., 2024). In some low resource code mixed language scenarios such as imbalanced classes and language variations like Kannada-English, Malayalam-English, Tamil-English, the models have performed consistently and competitively with other approaches like cost sensitive learning (Sreelakshmi et al., 2024).

3.5 **Advanced architectures and hybrid models**

Advanced architectures build a transformer and another component into one model, rather than combining separate models the way ensembles do. That extra piece could be a CNN, BiLSTM, attention layer, lexicon or graph network, added to catch something BERT misses on its own. Fazil et al. (2023) built one of these: several convolution channels pick up local patterns, then a two-layer BiLSTM with attention tracks the sequence, and it beat five leading models and five baselines, with the channels and attention doing most of the work.

Later work brings BERT into this same setup. Khan et al. (2022) built BiCHAT, which runs tweets through BERT, then an attention-based CNN, then an attention-based BiLSTM, and it beat earlier models on three Twitter datasets. Goswami and Daniel (2026) pair DistilBERT with a BiLSTM and tune it with Bayesian optimization. Ali et al. (2025) go a different way, adding a hate speech lexicon to BERT by flagging which words in the text show up in the lexicon, which helps catch misspellings and disguised hate. Awan et al. (2026) do the same for Roman Urdu with LEXF-ATT-XLM, combining XLM-RoBERTa, a BiGRU, attention and a Roman Urdu hate lexicon.

3.6 **LLM-based approaches (GPT, LLaMA)**

In recent years, much of the research has focused on general-purpose large language models (LLMs) for identifying hate speech, rather than the task-specific fine-tuning paradigm with

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

transformer encoders described earlier. Early work on GPT-3 showed it could detect sexist and racist text with no examples at all, and that adding a few labeled examples to the prompt raised accuracy to about 56 per cent (Chiu et al. 2021). This result of small improvements from zero-shot to few-shot prompting that are still quite significant, has been witnessed in later evaluations of performance between GPT-3.5, GPT-4 and open source models like LLaMA. In a benchmark on Assamese hate speech, for instance, the performance of GPT-4 (in the 5-shot setting) exceeded that of the fine-tuned multilingual BERT variants (mBERT), indicating that sufficiently large LLMs can outperform fine-tuned encoders in low-resource or code-mixed environments (Nath et al., 2026.).This isn't always true, though. In political science, a well-tuned BERT model far outperformed a prompted GPT model at text classification, and few-shot prompting barely improved on zero-shot (Wang et al., 2024). Other work found that smaller fine-tuned LLMs beat larger zero-shot ones (Fine-Tuned "Small" LLMs, 2025). Similarly, on the AfriHate corpus, the overall zero-shot F1 scores of GPT-4o were lower than those of fully fine-tuned models, though it scored the highest on the overall dataset (Muhammad et al., 2025). Overall, LLM performance depends on how much training data is available, how well the target language or domain is represented, and how that data is composed. Fine-tuned BERT models, or domain-adapted versions such as HateBERT (Caselli et al., 2021), tend to do better when the language or domain is underrepresented in the LLM's training, or when a large task-specific dataset is available. LLMs also have several common limitations, such as their inability to deal with code-mixed and transliterated text because they were not explicitly trained on such data (Rethinking Hate Speech Detection on Social Media, 2025), their weaker understanding of cultural nuances and sarcasm even with effective few-shot settings (Nath et al., 2026) and their sensitivity to prompt design and example selection, which makes few-shot deployment more costly (Bucher & Martini, 2024). LLaMA is a competitive and more open alternative to the closed GPT models for hate speech classification tasks, as seen in some benchmarks where it outperforms the proprietary models like GPT-4o-mini (Singh et al., 2025).

COMPARATIVE ANALYSIS

Here's the literature summarized as a table:

4.1 General BERT Fine-Tuning Approaches

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Liu et al.	2019	RoBERTa	GLUE, RACE, SQuAD	On the GLUE benchmark (dev set), it achieved F1 score of 92.2 on QQP and 90.9 on MRPC.	On the GLUE benchmark (dev set), RoBERTa achieved an accuracy of 90.2 on MNLI, 94.7 on QNLI, 86.6 on RTE, 96.4 on SST-2 and 91.3 on WNLI. On the SQuAD question-answering benchmark, RoBERTa reported F1 scores of 94.6 on SQuAD v1.1 and 89.4 on SQuAD v2.0 (dev set).	Refined BERT pretraining via hyperparameter and data-size tuning; showed BERT was undertrained; achieved SOTA on three benchmarks
Sanh et al.	2019	DistilBERT	GLUE,	The proposed model	DistilBERT	Introduced knowledge

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
			SQuAD and IMDb	gave F1 score of 85.8 on SQuAD v1.1 and with a second step of distillation during fine-tuning it improved to 86.9	reached a test accuracy of 92.82 on the IMDb sentiment classification benchmark	distillation during pre-training; 40% smaller, retains 97% of BERT's capability, 60% faster inference.
Rajput et al.	2021	Static BERT embeddings vs. fastText/GloVe	ETHOS	The best model in the paper that uses BiLSTM with static BE achieves F1 acore of 79.71%.	The best model in the paper that uses BiLSTM with static BE achieves accuracyof 80.15%.	Showed static BERT embeddings outperform fastText/GloVe; reduced false positives, improved specificity
Mozafari et al.	2019	Fine-tuned BERT (transfer learning)	Two Twitter datasets (racism, sexism, hate, offensive)	The proposed model achieved F1-score of 88% on Waseem dataset and 92% on Davidson dataset.	Not reported	Demonstrated improved precision/recall via transfer learning; showed BERT captures annotation/collection bias

3.2 Domain-Specific Transformer Models

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Caselli et al.	2021	HateBERT	Reddit corpus (banned communities) + 3 English abusive-language datasets	It achieved a macro-F1 of 80.9% on OffensEval 2019, 76.5% on AbusEval and 51.6% on HatEval dataset.	Not reported	Domain-specific retraining of BERT on abusive Reddit content; outperformed general BERT; tested cross-dataset portability
Li et al.	2021	COVID-HateBERT	COVID-19 related tweets	Not reported	Not reported	Showed general hate-speech models underperform on COVID-specific content; built thematically specialized model
Ocampo et al.	2023	Existing SOTA models -USE +SVM, BERT, HateBERT, DeBERTa V3	ISHate	DeBERTa V3 without augmentation reported Explicit HS F1as 85.1%, Implicit HS F1as 44.2%	Not reported	Highlighted difficulty of detecting implicit hate speech (circumlocution, metaphor, sarcasm)
Pérez et al.	2023	**BETO** (Spanish BERT-base) — with context (Tweet), fine-tuned	Contextualized Hate Speech Corpus (Rioplatense Spanish, COVID-19-related Twitter replies) d	The best model observed by authors is Tweet context, fine-tuned BETO which gave F1 score of 69.7%	Tweet context, fine-tuned BETO gave an accuracy of 91%	Identified lack of conversational context as a key limitation in hate speech detection

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

3.3 Ensemble Transformer Approaches

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Mnassri et al.	2022	Ensemble (stacking/voting of BERT-based models)	Combined public Twitter hate/offensive datasets (incl. Davidson)	**Stacking EM4** produced the F1 scores of 97% on Davidson, 92.78% on DHO dataset	**tacking EM4** produced accuracy of 77.6% on Davidson dataset and 46.3% on DHO dataset.	Applied multiple ensemble strategies (soft/hard/max voting, stacking); stacking most effective
Siddiqui et al.	2024	Ensemble of multilingual transformers (XLM-RoBERTa) + explainable AI	English, Urdu, Sindhi datasets	Up to 91% (binary); 86% weighted (fine-grained)	Not reported	Combined multilingual ensembling with explainability for fine-grained multilingual classification
Yoo et al.	2025	Parallel ensemble of BERT-based models(BERT, DistilBERT and RoBERTa)	English and Korean hate-speech datasets	85% (English); 89% (Korean)	85% (English); 89% (Korean)	Used majority voting/weighted averaging with a meta-learner; outperformed single-model baselines

3.4 Multilingual Transformer Approaches

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Mozafari et al.	2022	Proto-MAML (meta-learning)	15 datasets, across 8 languages (hate speech); 6 datasets across 6 languages (offensive language) — HatEval, HASOC, OffensEval, EVALITA, etc.	Proto-MAML, k=16, hate speech task gave average macro F1 of 62.68%	Not reported	Enabled rapid cross-lingual few-shot adaptation; outperformed conventional transfer learning.
Miok et al.	2022	Multilingual BERT + Bayesian attention (Monte Carlo dropout)	English tweets (subset of Davidson et al.'s dataset), Croatian news comments (comments from the Večernji list news portal) and Slovene Facebook comments (from the FRENK project).	MCD BERT provided the best result across all three languages (English, Croatian, Slovene) and gave F1 score of 90.4% in English	MCD BERT provided the best result across all three languages (English, Croatian, Slovene) and gave an accuracy of 91.4% for English tweets.	Added calibrated uncertainty estimation for trustworthy content-moderation predictions

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Madhavi et al.	2024	A hybrid CNN-GRU + BERT architecture	The Davidson et al. hate speech and offensive language dataset	Proposed CNN-GRU-BERT gave F1 score of 96%	Proposed CNN-GRU-BERT gave the accuracy of 98%	Combined contextual and sequential features across languages
Sreelakshmi et al.	2024	MuRIL + cost-sensitive learning	Dravidian code-mixed (Kannada-English, Malayalam-English, Tamil-English)	The best F1 score out of six datasets tested is 96% and it was given by DistilBERT + SVM RBF on DravidianLangTech 2021 Malayalam dataset.	The best accuracy out of six datasets tested is 96.2% and it was given by DistilBERT + SVM RBF on DravidianLangTech 2021 Malayalam dataset.	Addressed class imbalance and linguistic variability in low-resource code-mixed settings.

3.5 Advanced Architectures and Hybrid Models

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Fazil et al.	2023	Attentional multi-channel CNN-BiLSTM	Three Twitter benchmark datasets	On the imbalanced version of DS-2 (Davidson et al.), the proposed model achieved F1 score of 92.79%	On the imbalanced version of DS-2 (Davidson et al.), the proposed model achieved an accuracy of 92.48%	Combined multi-kernel CNN channels with attention-based BiLSTM; outperformed 10 baseline/SOTA models
Khan et al.	2022	BiCHAT (BERT + attention-aware CNN + attention-aware BiLSTM)	Three Twitter benchmark datasets	Outperformed SOTA and baselines by 8% on f-score (abstract reports relative gains, not absolute F1)	Outperformed SOTA and baselines by 5% (training) and 9% (validation) accuracy	Combined BERT with a hierarchical attention CNN and BiLSTM; outperformed prior state-of-the-art and baseline models
Goswami & Daniel	2026	DistilBERT + BiLSTM (Bayesian-optimized)	Jigsaw Toxic Comment Classification	94%	95%	Combined DistilBERT with BiLSTM and used Bayesian optimization to tune hyperparameters
Ali et a	2025	BERT + lexicon-based features (hybrid)	100,000 social media posts + Founta et al. (2018)	Not reported	Up to 95%	Added an adaptive, updatable hate lexicon to BERT to catch evolving slurs, misspellings and

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
			dataset			coded language
Awan et al.	2026	LEXF-ATT-XLM (XLM-RoBERTa + BiGRU + attention + lexicon)	RU-HSD-30K (Roman Urdu)	88.83% (weighted F1)	88.83%	Added a Roman Urdu hate lexicon as an auxiliary supervision signal alongside multilingual transformer embeddings

3.6 LLM-Based Approaches (GPT, LLaMA)

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
Chiu et al.	2021	GPT-3 (zero-/few-shot)	Comments and text extracts labeled for sexist and racist content from ETHOS dataset.	Under the mixed-category few-shot learning setting, the proposed model gave F1 score of 79% for racist comments and 77% for sexist comments.	Under the mixed-category few-shot learning setting, the proposed model gave accuracy of 83% for racist comments and 85% for sexist comments.	It is testing a large language model via prompting rather than training a classifier.it is an early evaluation of GPT-3 for sexist/racist text and showed few-shot prompting improves performance.
Nath et al.	2026	GPT-4 (five-shot) vs. mBERT	Assamese hate speech dataset	GPT-4 achieved the best performance in the few-shot (k=5) configuration and gave F1 score of 87%	GPT-4 achieved the best performance in the few-shot (k=5) configuration and achieved accuracy of 88.5 %	GPT-4 (five-shot) outperformed fine-tuned mBERT in a low-resource, code-mixed setting
Wang et al.	2024	Fine-tuned BERT vs. prompted GPT	Political science text classification dataset	Not specified	Not specified	Found fine-tuned BERT substantially outperformed GPT; few-shot gave little benefit over zero-shot
Bucher & Martini	2024	Small fine-tuned LLMs vs. large zero-shot generative models	Not specified	Not specified	Not specified	Found smaller fine-tuned LLMs consistently outperformed larger zero-shot models
Muhammad et al.	2025	GPT-4o vs. fine-tuned models	AfriHate corpus	Not specified	Led in aggregate zero-shot F1	GPT-4o led in aggregate F1 but underperformed fine-tuned models on a per-class basis
Singh et al.	2025	LLaMA-based models vs. GPT-4o-mini	HateXplain, ImplicitHate, IndoHateMix	Fine-tuned LLaMA-3.1-8B achieved the highest F1 across all three datasets 0.91 on HateXplain, 0.81 on ImplicitHate, 0.82 on IndoHateMix.	Fine-tuned LLaMA-3.1-8B achieved the highest accuracy across all three datasets 0.91on HateXplain, 0.82 on ImplicitHate, 0.83on IndoHateMix), outperforming	Open-weight LLaMA models outperformed GPT-4o-mini in some benchmarks; noted weaknesses on code-mixed/transliterated text

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

Paper	Year	Model	Dataset	F1 Score	Accuracy	Key Contribution
					fine-tuned mBERT on every benchmark.	

As a whole, these papers do not suggest a definite answer as to which of the two types of models— generative LLMs or fine-tuned BERT models— is superior for text classification. Both studies, Muhammad et al. (2025) and Bucher & Martini (2024), conclude that finetuned encoder models outperform zero-shot and few-shot LLMs prompting, with difference being greatest for more challenging, low-resource or more specific tasks. In a more conditional finding, Wang et al. (2024) found that GPT-4o can be as effective as fine-tuned RoBERTa when the tasks are limited to binary classification, but fine-tuning outperforms GPT-4o when it comes to multclass classification. While Muhammad et al. (2025) and Bucher and Martini(2024) found that a fine-tuned BERT-based model is superior, Singh et al. (2025) are able to show that a fine-tuned LLaMA-3.1-8B model achieves better performance than the fine-tuned BERT models, even in low-resource code-mixed settings. In this context, it seems that while a smaller encoder model may be preferable for some tasks, the decision between the two approaches is more likely to be driven by the difficulty of the task, the number of classes, language resource availability and the type of LLM used (fine-tuned vs. untrained) than by any general model superiority.

CHALLENGES AND OPEN ISSUES

5.1 Class imbalance

Most of the datasets used in the papers reviewed here lean heavily toward the non-hate class. Davidson et al. 's dataset, for instance, has hateful posts making up only a small fraction of the total and even ETHOS, which was built specifically for hate speech research, is not evenly split. When a model is trained on data like this, it can reach a high overall accuracy simply by leaning toward the majority class, while its recall on the minority (hateful) class suffers. That is a serious problem for a moderation system, since missing actual hate speech is a costlier mistake than flagging a borderline post. Several of the papers surveyed here address this directly. The idea of cost sensitive learning (Sreelakshmi et al., 2024) is to impose higher penalty on the misclassification of the minority class and it has been demonstrated that stacking classifiers (Mnassri et al., 2022) can also reduce the impact of skewed class distribution. Oversampling, class weighted loss functions as well as the careful selection of the evaluation metric (F1 or macro-F1 instead of raw accuracy) all appear as partial remedies but they don't completely address the issue of the lack of data for the hateful class.

5.2 Lack of explainability

Transformer models are effectively black boxes: they produce a label, but not a reason. For content moderation, the lack of explainability is a real limitation as a human reviewer has to decide whether a flagged post is actually hateful or not. In the worst case, one wrongly removed post can generate legitimate complaints from users. The model's reasoning is not accessible to the moderator who uses only a raw "hate" or "not hate" tag. He can't catch a mistake in time before it will impact someone's account. To fill this void, Hashir et al (2025) propose a novel framework termed as TARGE. It is a model that produces rationales in natural language, in addition to the label itself, for a classification decision, using an LLM. This kind of explicit, human-readable explanation is a more promising direction than earlier explainability work, since it gives moderators something to actually reason about rather than a heatmap over tokens. Still, most of the models covered elsewhere in this review, including

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

the domain-specific and ensemble approaches from Sections 3.2 and 3.3, offer no explanation at all, which remains one of the more practical barriers to deploying these systems in real moderation pipelines.

5.3 Sarcasm and context

Most of the hate speech doesn't use slurs or sound aggressive. It hides behind sarcasm, code words, or references that only make sense if you know who's talking, who they're talking to and what they mean. As found in the study by Ocampo et al. (2023), indirect, metaphorical and ironic hate messages are far more difficult to detect than explicit hate messages, even for a robust model such as DeBERTa V3, where the F1 score was significantly lower in the implicit cases. However, Pérez et al. (2023) highlight another challenge: an individual tweet/comment on its own is often not sufficiently informative to help determine if it is hateful and models that do not take into account the surrounding conversation miss information that a human reader would easily notice. Their work on Rioplatense Spanish Twitter replies found that feeding tweet context into a fine-tuned BETO model improved results over judging each post in isolation. Chapagain et al.'s work on the MetaHate collection reaches a similar conclusion from a different angle, reporting that sarcasm and coded language remain among the biggest sources of classification errors even for their best-performing model. All of this indicates that most real-world hate speech is not targeted to any single message, as is the case with most datasets and models used in this review, but rather remains undetected.

5.4 Multilingual and code-mixed text

The majority of the datasets and models mentioned in this review are developed for English, such as Davidson et al., ETHOS and HateXplain. Real social media usage does not involve just English usage and a significant amount of this usage is code-mixed—one post using two or more languages in the same sentence, sometimes in the same word. A similar pattern appears repeatedly across all South Asian and other multilingual online communities, but English only models are not designed to process such input. The multilingual work in Section 3.4 is a step towards bridging this gap: Sreelakshmi et al. (2024) demonstrates that MuRIL combined with cost-sensitive learning works well on Kannada-English, Malayalam-English and Tamil-English code-mixed datasets, while Mozafari et al. (2022) shows that a model trained on a source language can be meta-learned to generalise to a new target language with just a few labelled examples without having to run a full retraining session. Despite this, a larger number of these are still minority studies compared to what is available in English-language studies and even with the languages surveyed here, there is still insufficient coverage across the languages of the world, particularly the low-resource languages.

5.5 Dataset limitations

Beyond class imbalance, dataset size poses a further limitation. ETHOS (Mollas et al., 2022), one of the benchmarks used in this review, has fewer than 1,000 annotated comments. That’s a small sample for training or evaluating a deep model with any confidence, especially given how skewed the split already is between hateful and non-hateful comments. There are larger resources, but they come with their own problems. MetaHate (Chapagain et al., 2025) solves the size problem by combining 36 datasets into one large collection of 1.2 million samples. But the annotation guidelines aren't consistent across the source datasets, since one may define "hateful" differently from another. Such inconsistencies make comparisons between results of papers more difficult as the reported accuracy or F1 score of a model is only meaningful if they are compared to a set of labels to which the model was trained and tested. Of course, building larger datasets matters, but so does agreeing on clearer annotation

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

guidelines. The field also needs more coverage of different languages, dialects and platforms, since most current research focuses on tweets.

FUTURE DIRECTIONS

Building on the gaps identified above, a few directions seem the most likely to move this field forward.

1. Fine-tuning open-weight LLMs on augmented low-resource data.

Singh et al. (2025) already showed that a fine-tuned LLaMA-3.1-8B model beats fine-tuned BERT-like models on code-mixed datasets. The same strategy might work well on smaller, harder datasets like ETHOS, without needing a custom model like GPT-4.

2. Building explainability from the start

Instead of adding attention visualization to an already-trained classifier, some approaches generate a rationale alongside the label. Hashir et al. 's (2025) TARGE framework is one example, and this direction seems more useful for models that moderators actually have to act on.

3. Multimodal detection

The topic of hateful content on social media is rarely just about text; memes and images often have as much signal as the text and user metadata (posting history, network connections) are a logical next step beyond the text-only models covered in this review.

4. Wider coverage of low-resource and code-mixed languages

Meta-learning and cost-sensitive methods, discussed in Section 3.4, are promising but only cover a few language pairs. Expanding them to more languages and dialects would fill a real gap.

5. Real-time deployment

Most models in this review are tested offline on static benchmarks, but a real moderation system needs to classify content within seconds of it being posted. Distillation approaches like DistilBERT (Sanh et al., 2019) show how to balance accuracy and speed for this kind of deployment.

CONCLUSION

This paper maps how transformer-based models and large language models are being applied for the purpose of hate speech detection. It identifies which approaches actually work and where they still fall short. Across the studies covered here, transformer-based models consistently outperform the older CNN, LSTM and traditional machine learning baselines. Domain-specific or ensemble transformer approaches also tend to outperform single general-purpose models within that group. HateBERT and similar domain-retrained encoders, along with stacking ensembles like Mnassri et al.'s (2022), are some of the strongest performers reviewed here. Fine-tuned LLaMA models are starting to close that gap with fine-tuned BERT variants, and in some low-resource settings, even reverse it. The most pressing open problem is still limited explainability combined with weak coverage of non-English and code-mixed language. Both stand in the way of deploying these models safely and reliably in real moderation systems. As LLMs keep improving and explainability techniques like Hashir et al.'s (2025) mature, hate speech detection is likely to become more accurate, more transparent, and better suited to the multilingual reality of social media. The ensemble of multiple models in combination with BERT model and other classical machine learning

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

based classifiers or other well known deep learning architectures generally outperform the other models in the Hate Speech Detection and Classification tasks.

REFERENCES

1. Alrehili, A. (2019, November). Automatic hate speech detection on social media: A brief survey. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-6). IEEE.
2. Ali, S., Blackburn, J., & Stringhini, G. (2025). Evolving hate speech online: An adaptive framework for detection and mitigation. arXiv:2502.10921.
3. Awan, J. J., Hamid, M., & Alshalali, T. A. N. (2026). LEXF-ATT-XLM: A hybrid lexicon-enhanced attention model for hate speech detection in low-resource language Roman Urdu. PLoS One, 21(8), e0354875.
4. Bilal, M., Khan, A., Jan, S., Musa, S., & Ali, S. (2023). Roman Urdu hate speech detection using transformer-based model for cyber security applications. Sensors, 23(8), 3909.
5. Brison, S. (2013). Hate speech. The International Encyclopedia of Ethics, 2332-2342.
6. Bucher, M. J. J., & Martini, M. (2024). Fine-tuned "small" LLMs (still) significantly outperform zero-shot generative AI models in text classification. arXiv preprint arXiv:2406.08660.
7. BiCHAT: Khan, S., Fazil, M., Sejwal, V. K., Alshara, M. A., Alotaibi, R. M., Kamal, A., & Baig, A. R. (2022). Journal of King Saud University – Computer and Information Sciences, 34(7), 4335–4344.
8. Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 17-25). Association for Computational Linguistics.
9. Chapagain, S., Hamdi, S. M., & Boubrahimi, S. F. (2025, August). Advancing hate speech detection with transformers: Insights from the metahate. In International Conference on Advances in Social Networks Analysis and Mining (pp. 432-439). Cham: Springer Nature Switzerland.
10. Chiu, K.-L., Collins, A., & Alexander, R. (2021). Detecting hate speech with GPT-3. arXiv preprint arXiv:2103.12407.
11. Desai, P. N., Kewalramani, T., & Mandal, S. (2025). Feature Selection Empowered BERT for Detection of Hate Speech with Vocabulary Augmentation. arXiv preprint arXiv:2512.02141.
12. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
13. Fazil, M., Khan, S., Albahlal, B. M., Alotaibi, R. M., Siddiqui, T., & Shah, M. A. (2023). Attentional multi-channel convolution with bidirectional LSTM cell toward hate speech prediction. IEEE Access, 11, 16801-16811.
14. Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. Unesco Publishing.
15. Gelber, K. (2021). Differentiating hate speech: A systemic discrimination approach. Critical Review of International Social and Political Philosophy.
16. Ghosh, K., & Senapati, A. (2022, October). Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (pp. 853-865).
17. Goswami, P., & Daniel, A. (2026). Enhancing Hate Speech Detection with a Distil BERT and BiLSTM Hybrid Mode. SN Computer Science, 7(4), 353.
18. Hashir, M. H., & Kim, S. W. (2025). TARGE: large language model-powered explainable hate speech detection. PeerJ Computer Science, 11, e2911.
19. Kalra, S., Sanchit, Gupta, V., Jha, S., & Shukla, A. (2025, December). Context-Aware Hate and Offensive Speech Detection Using Hybrid BERT-LSTM Architecture. In International Conference on Information and Communication Technology for Competitive Strategies (pp. 219-234). Cham: Springer Nature Switzerland.

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

20. Li, M., Liao, S., Okpala, E., Tong, M., Costello, M., Cheng, L., ... & Luo, F. (2021, December). COVID-HateBERT: A pre-trained language model for COVID-19 related hate speech detection. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 233-238). IEEE.
21. Lexicon hybrid: Ali, S., Blackburn, J., & Stringhini, G. (2025). "Evolving Hate Speech Online: An Adaptive Framework for Detection and Mitigation." arXiv:2502.10921.
22. Madhavi, M., Agal, S., Odedra, N. D., Chowdhary, H., Ruprah, T. S., Vuyyuru, V. A., & El-Ebiary, Y. A. B. (2024). Elevating offensive language detection: CNN-GRU and BERT for enhanced hate speech identification. International Journal of Advanced Computer Science & Applications, 15(5).
23. Miok, K., Škrlj, B., Zaharie, D., & Robnik-Šikonja, M. (2022). To ban or not to ban: Bayesian attention networks for reliable hate speech detection. Cognitive Computation, 14(1), 353-371. —
24. Mittal, U. (2023, August). Detecting hate speech utilizing deep convolutional network and transformer models. In 2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM) (pp. 1-4). IEEE.
25. Mnassri, K., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2022, December). BERT-based ensemble approaches for hate speech detection. In GLOBECOM 2022-2022 IEEE global communications conference (pp. 4649-4654). IEEE.
26. Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: a multi-label hate speech detection dataset. Complex & Intelligent Systems, 8(6), 4663-4678.
27. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In International Conference on Complex Networks and Their Applications (pp. 928-940). Springer.
28. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. IEEE Access, 10, 14880-14896.
29. Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Adelani, D. I., Ahmad, I. S., Aliyu, S. M., ... & Ousidhoum, N. (2025, April). AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 1854-1871).
30. Nath, B., Pandey, K. K., Yadav, V., & Chintalapati, P. V. (2026). Few-shot and zero-shot Assamese hate speech detection: A comparative benchmark of large language models. Social Network Analysis and Mining, 16(1), 14.
31. Ocampo, N. B., Sviridova, E., Cabrio, E., & Villata, S. (2023, May). An in-depth analysis of implicit and subtle hate speech messages. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 1997-2013).
32. Pérez, J. M., Luque, F. M., Zayat, D., Kondratzky, M., Moro, A., Serrati, P. S., ... & Cotik, V. (2023). Assessing the impact of contextual information in hate speech detection. IEEE Access, 11, 30575-30590.
33. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection. Language Resources and Evaluation, 55(2), 477-523.
34. Rajput, G., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2021, December). Hate speech detection using static bert embeddings. In International conference on big data analytics (pp. 67-77). Cham: Springer International Publishing.
35. Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., ... & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. Social Network Analysis and Mining, 14(1), 204.
36. Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1-10).

© Copyright @ Authors

ISSN: 2350-1278
Peer Reviewed & Refereed Journal (IF: 7.9)
Journal Website www.nrjitis.in

37. Sellars, A. (2016). Defining hate speech. Berkman Klein Center Research Publication, (2016-20), 16-48.

38. Shawkat, N., Saquer, J., & Shatnawi, H. (2024, April). Evaluation of different machine learning and deep learning techniques for hate speech detection. In Proceedings of the 2024 ACM Southeast Conference (pp. 253-258).

39. Siddiqui, J. A., Yuhaniz, S. S., Shaikh, G. M., Soomro, S. A., & Mahar, Z. A. (2024). Fine-grained multilingual hate speech detection using explainable AI and transformers. IEEE Access, 12, 143177-143192.

40. Singh, D. D., Bhattacharjee, R., & Chakraborty, A. (2025). Rethinking hate speech detection on social media: Can LLMs replace traditional models? arXiv preprint arXiv:2506.12744.

41. Sreelakshmi, K., Premjith, B., Chakravarthi, B. R., & Soman, K. P. (2024). Detection of hate speech and offensive language codemix text in Dravidian languages using cost-sensitive learning approach. IEEE Access, 12, 20064-20090.

42. Tiţa, T., & Zubiaga, A. (2021). Cross-lingual hate speech detection using transformer models. arXiv preprint arXiv:2111.00981.

43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

44. Vergani, M., Perry, B., Freilich, J., Chermak, S., Scrivens, R., Link, R., ... & Iqbal, M. (2024). Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech and hate incidents: A systematic review. Campbell Systematic Reviews, 20(2), e1397.

45. Vilar-Lluch, S. (2023). Understanding and appraising 'hate speech'. Journal of Language Aggression and Conflict, 11(2), 279-306.

46. Wang, Y., Qu, W., & Ye, X. (2024). Selecting between BERT and GPT for text classification in political science research. arXiv preprint arXiv:2411.05050.

47. Yoo, S., Jeon, E., Hyeon, J., & Cho, J. (2025). Adaptive ensemble techniques leveraging BERT based models for multilingual hate speech detection in Korean and english. Scientific Reports, 15(1), 19844.

© Copyright @ Authors