



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 10, October 2026

A Survey on Classification of Hate Speech using Machine Learning

Khierod Behera¹, Prof. Megharaj Patil², Sanmesh Bhalerao³, Ramesh Shinde⁴, Vishal Shinde⁵

Department of Information Technology, Genba Sopanrao Moze College of Engineering, Pune, India¹⁻⁵

ABSTRACT: Hate speech detection is an important task in natural language processing, particularly for low-resource languages where linguistic resources and annotated datasets are limited. This study presents a comparative analysis of Large Language Models (LLMs) for detecting and classifying hate speech. The study evaluates five LLMs: LLaMA 3, Qwen3, DeepSeek-R1, Mistral, and Phi-3. Different prompting strategies, including single-stage sentence-level, two-stage sentence-level, and paragraph-level classification, are examined. The models are evaluated on a twitter dataset containing non-hate, offensive, and hate speech categories. The experiments investigate the effectiveness of zero-shot and few-shot prompting approaches and ensemble majority voting.

KEYWORDS: Hate Speech Detection, Large Language Models, Natural Language Processing, Low-Resource Languages, Text Classification

I. INTRODUCTION

Hate speech detection has become an important research area in Natural Language Processing (NLP) due to the rapid growth of social media and online communication. The increasing amount of user-generated content has created a need for automated systems that can identify and classify harmful language. Hate speech detection involves distinguishing between non-hate, offensive, and hateful expressions based on their textual content. However, identifying hate speech is challenging because words and expressions can have different meanings depending on their context. Accurate classification is important for understanding harmful online content and supporting content moderation.

Detecting hate speech in low-resource languages is particularly challenging because of the limited availability of annotated datasets, linguistic resources, and language-specific computational tools. Traditional machine learning approaches often require sufficient labelled training data and carefully designed language-specific features. Limited resources can affect their ability to recognize different forms of offensive and hateful expressions. Therefore, evaluating methods that can classify hate speech with fewer language-specific resources is an important research direction.

Large Language Models (LLMs) have introduced new approaches to text classification through natural language instructions and prompting techniques. This study presents a comparative analysis of five LLMs: LLaMA 3, Qwen3, DeepSeek-R1, Mistral, and Phi-3, for hate speech detection. It investigates single-stage sentence-level, two-stage sentence-level, and paragraph-level prompting strategies to examine their classification performance. The study also considers zero-shot and few-shot prompting and ensemble majority voting. The models are evaluated on three categories: non-hate, offensive, and hate speech. This comparative analysis examines the effectiveness of different models and prompting strategies for hate speech classification in a low-resource language.

II. LITERATURE REVIEW

Dibyanayan Bandyopadhyay et al. [1] – The study proposes M-NLE, a knowledge-augmented multitask learning model for offensive meme detection and explanation generation. The model combines external knowledge with multitask learning to identify offensive content in memes and generate explanations for its predictions. It aims to improve the understanding of offensive memes by considering both detection and explanation tasks. This approach also focuses on making meme classification more interpretable.

Haifa Saleh Alfurayj et al. [2] – The study introduces CYBY24 and a step-wise model for thread-based, fine-grained cyberbullying detection. The proposed approach analyzes online discussion threads to identify cyberbullying by considering the relationships and context between messages. It focuses on detailed classification rather than simply identifying whether cyberbullying is present. This approach helps address the challenges of detecting harmful language in complex online conversations.

EbtihalEnfes et al. [3] – The study presents a cross-dialectal transfer learning approach for offensive language detection in Arabic. The proposed method uses knowledge transferred from one Arabic dialect to another to address linguistic differences and variations in offensive expressions. It focuses on improving classification across multiple dialects, particularly where labelled training data is limited. This approach explores the use of transfer learning to support offensive language detection in multilingual and low-resource environments.

Nathanael P. Solagratia et al. [4] – The study investigates the use of BERT and local vision-language models for hate speech detection on small multimodal datasets. It explores how textual and visual information can be combined to identify hateful content in online media. BERT is used to process textual information, while vision-language models help analyze visual content. The research focuses on applying these models to datasets with limited samples and examines the challenges of multimodal hate speech detection.

Avital Rozenthal et al. [5] – The study presents a systematic analysis of oversampling techniques and traditional machine learning pipelines for multilingual offensive language detection in resource-constrained environments. It focuses on addressing class imbalance, where offensive and non-offensive language samples may not be equally distributed. The research examines how oversampling techniques can support classification across different languages while considering computational limitations. It also explores the applicability of traditional machine learning methods for multilingual offensive language detection when resources are limited.

Maximilian Nienhaus et al. [6] – The study explores the use of large language models for German hate speech detection by leveraging conversational context. It focuses on analyzing surrounding messages to understand the meaning and intention of potentially hateful expressions. The approach considers contextual information to improve the identification of hate speech in online conversations. It also addresses the challenges of detecting hate speech when individual messages lack sufficient context.

Mahmoud Abusaqer et al. [7] – The study proposes BERT-OTA, an ontology-guided transformer attention model for hate speech detection. The approach integrates domain-specific ontological knowledge with transformer-based attention to improve the identification of hateful expressions. It focuses on combining linguistic features with structured knowledge to support more context-aware classification. This method aims to improve the representation of hate-related terms and their relationships during the detection process.

Farisha Azmi et al. [8] – The study presents a systematic review and research taxonomy of culturally aware Malay–English code-mixed hate speech detection. It focuses on the challenges of identifying hate speech in social media content that combines Malay and English languages. The research examines linguistic variations, cultural context, and code-mixing characteristics that affect classification. It also organizes existing research approaches and identifies potential research gaps for developing more effective hate speech detection systems.

Kathiravan Pannerselvam et al. [9] – The study introduces IRAIVI, an agentic framework for Tamil hate speech detection and counter-speech generation. The framework focuses on identifying hateful content in Tamil and generating appropriate counter-speech responses. It explores the use of agentic techniques to support hate speech analysis and responsible intervention. The approach also addresses the importance of language-specific understanding and generating responses that can help reduce harmful online communication.

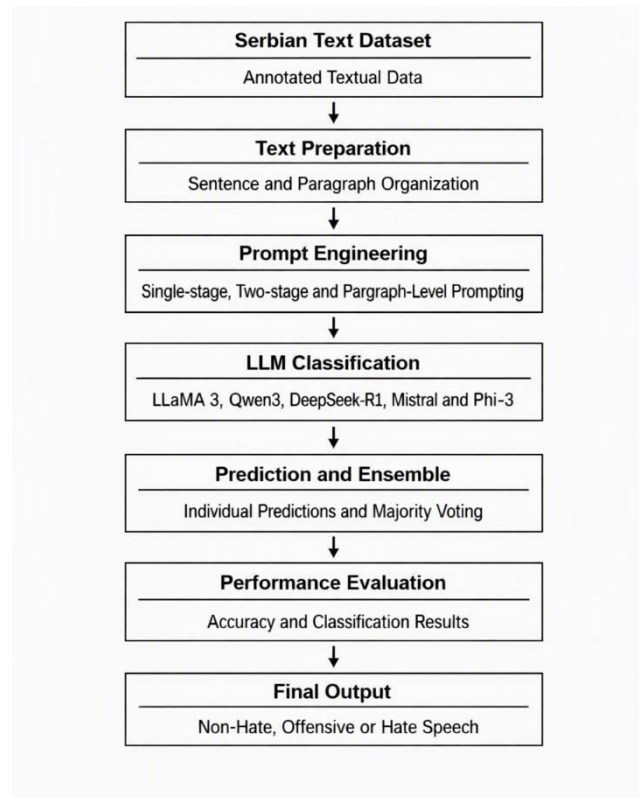
Sr. No	Paper Name	Advantages	Limitation
1.	M-NLE: Knowledge-Augmented Multitask Learning for Offensive Meme Detection and Explanation Generation	Improves meme detection and provides explanations.	Requires additional knowledge and computational resources.
2.	CYBY24 and Step-Wise Model for Thread-Based Fine-Grained Cyberbullying Detection	Considers thread context and fine-grained cyberbullying categories.	Depends on thread context and labelled data.
3.	Cross-Dialectal Transfer Learning for Offensive Language Detection in Arabic	Supports offensive language detection across Arabic dialects.	Dialect variations may affect detection performance.
4.	BERT and Local Vision Language Models in Hate Speech Detection on Small Multimodal Datasets	Combines text and visual information for hate speech detection.	Limited datasets may affect generalization.
5.	A Systematic Analysis of Oversampling and Classic Machine Learning Pipelines for Multilingual Offensive Language Detection in Resource-Constrained Scenarios	Addresses class imbalance with computationally accessible methods.	Traditional models may struggle with complex language context.

6.	Leveraging Conversational Context for German Hate Speech Detection With Large Language Models	Uses conversational context to identify hate speech more effectively.	Requires sufficient contextual information and computational resources.
7.	BERT-OTA: Enhancing Hate Speech Detection With Ontology-Guided Transformer Attention	Integrates ontological knowledge with transformer attention.	Depends on the quality and coverage of the ontology.
8.	Culturally Aware Malay–English Code-Mixed Hate Speech Detection: A Systematic Review and Research Taxonomy	Identifies challenges in culturally aware code-mixed hate speech detection.	Review-based study without a single proposed detection model.
9.	IRAI VI: Agentic Framework for Tamil Hate Speech Detection and Counter-Speech Generation Toward Responsible AI Intervention	Combines Tamil hate speech detection with counter-speech generation.	Requires language-specific data and reliable response generation.

III. RESEARCH GAP

- Limited language resources: Low-resource languages have limited annotated datasets for hate speech detection.
- Limited comparative evaluation: There is a need to compare different LLMs using a common dataset and evaluation setup.
- Prompting strategy limitations: The effect of sentence-level, two-stage, and paragraph-level prompting requires further investigation.
- Classification challenges: Distinguishing between non-hate, offensive, and hate speech remains challenging because of linguistic and contextual differences.
- Limited ensemble evaluation: The effectiveness of combining predictions from different LLMs through majority voting needs further investigation.

IV. ARCHITECTURE



V. METHODOLOGY

The proposed methodology focuses on the comparative evaluation of Large Language Models (LLMs) for detecting and classifying hate speech in the english language. It involves dataset preparation, prompt engineering, model selection, text classification, ensemble majority voting, and performance evaluation. The methodology examines how different models and prompting strategies affect the classification of non-hate, offensive, and hate speech.

1. Dataset Collection and Preparation

The study uses a twitter hate speech dataset containing 1,351 short texts and 8,029 sentences. The dataset consists of three classification categories: non-hate, offensive, and hate speech. These categories provide the reference labels for evaluating the classification results of the LLMs. The dataset is used to investigate how effectively different models recognize hateful and offensive expressions in a low-resource language.

2. Text Input and Organization

The collected twitter text is organized into suitable input formats for the experiments. The study considers sentence-level and paragraph-level classification. In sentence-level classification, individual sentences are provided to the models, whereas paragraph-level classification uses longer text units. These formats allow the study to examine how the amount of textual context affects the classification process. The input is supplied to the models through carefully designed prompts.

3. Prompt Engineering

Prompt engineering is used to guide the LLMs in identifying and classifying hate speech. Different prompting strategies are evaluated to understand their effects on classification performance.

- Single-stage sentence-level prompting: The model receives an individual sentence and directly predicts its category. The model classifies the sentence as non-hate, offensive, or hate speech in one step.
- Two-stage sentence-level prompting: The classification process is divided into two stages. The first stage identifies relevant hate-related content, while the second stage performs the required classification. This approach is used to investigate whether separating the classification process affects prediction accuracy.
- Paragraph-level prompting: The model receives a paragraph instead of an individual sentence. It analyzes the provided text and produces the required classification, allowing the evaluation of a broader textual context.

4. Two-Stage Sentence-Level Classification

The two-stage approach separates hate speech detection from fine-grained classification. Each sentence is processed independently, and the model performs the following operations.

- Stage 1 – Hate Speech Detection: The model receives a sentence and determines whether it contains hate speech. It generates a binary response, YES or NO, along with a short explanation. Offensive language without hate speech is treated as negative in this detection task.
- Stage 2 – Fine-Grained Classification: If the first stage detects hate speech, the same sentence is passed to a second prompt. The model assigns the relevant category and subcategory codes. Multiple codes are allowed when a sentence contains more than one type of hate speech.

The detection output is represented by `has_hs`, while the category output is represented by `class_id`. If the detection result is negative, the classification stage is skipped. This approach allows the detection and categorization stages to be evaluated separately.

5. Single-Stage Sentence-Level Classification

The single-stage approach performs hate speech detection and classification together using one prompt. Each sentence is provided to the model, which determines whether hate speech is present and assigns the corresponding category and subcategory codes when applicable.

The model is instructed to return exactly two outputs: `has_hs` and `class_id`. If hate speech is absent, the class ID is set to zero. If hate speech is detected, the model provides the relevant category codes. When multiple categories apply, the model can return multiple codes separated by commas.

This approach reduces the need for separate model calls and evaluates whether joint detection and classification can provide consistent results. Its performance is compared with the two-stage approach to examine the effect of task separation.

6. Paragraph-Level Classification

Paragraph-level classification is used to examine the effect of broader textual context on hate speech detection. Each input sample consists of a paragraph containing multiple sentences. The model is instructed to divide the paragraph into individual sentences without changing their original content and classify each sentence separately.

For every sentence, the model identifies whether hate speech is present and assigns the corresponding category and subcategory, if applicable. The output contains a sentence ID, a hate speech indicator, a category ID, and a subcategory ID. This method allows the model to consider surrounding sentences when identifying implicit, referential, or context-dependent hate speech.

The paragraph-level approach is particularly useful for studying cases in which the meaning of a sentence depends on the surrounding discussion. It also allows the evaluation of classification performance when a paragraph contains a mixture of neutral, offensive, and hateful sentences.

7. Ensemble-Based Classification

The ensemble approach combines predictions from multiple LLMs to investigate whether collective predictions improve classification robustness. The five selected models are executed concurrently using multithreaded inference. Their predictions for hate speech detection and category assignment are collected and compared.

Majority voting is applied to the model predictions to obtain the final decision. For binary detection, the system uses the majority decision on whether hate speech is present. Category predictions are also combined through voting. This approach examines whether combining different models can reduce individual model errors and improve the consistency of classification.

However, running multiple models for each input increases computational requirements and inference latency. Therefore, the ensemble results are also considered in relation to the additional computational cost.

8. Few-Shot Prompting

The study extends the two-stage sentence-level approach with few-shot prompting. In this setup, the first-stage detection prompt remains unchanged and uses zero-shot prompting. The second-stage classification prompt is supplemented with labelled demonstration examples.

The examples cover different classification situations, including non-hate sentences, offensive language, single-category hate speech, multiple-category hate speech, and borderline cases. These examples guide the model in assigning fine-grained labels and following the required output format. This setup is used to examine whether providing examples during categorization improves classification performance and how the effects differ across models.

9. Model Execution and Experimental Configuration

All five LLMs are hosted locally using Ollama and accessed through its REST interface. The experiments use identical prompts, input preprocessing, and specified decoding settings to maintain consistency between models. The temperature is set to 0.1, the generation limit is 1,024 tokens, and a fixed generation seed is used for the main experiments.

The study relies on instruction-based inference rather than task-specific fine-tuning for the main LLM experiments. The same category definitions and classification instructions are provided to the models so that their predictions can be compared under consistent conditions.

10. Performance Evaluation and Robustness Testing

The predictions generated by the models are compared with the human-annotated labels. The study evaluates binary hate speech detection, category-level classification, and fine-grained subcategory classification. Accuracy, precision, recall, F1-score, and inference latency are considered in the comparative evaluation.

In addition, the study uses five fixed generation seeds to examine the stability of zero-shot results. Bootstrap confidence intervals and McNemar's tests are also used for statistical analysis. These experiments help investigate whether observed differences between models and prompting strategies remain consistent across repeated runs.

Conclusion

This study presents a comparative analysis of five Large Language Models for hate speech detection and classification in the low-resource language. It examines different prompting strategies, including single-stage sentence-level, two-stage sentence-level, and paragraph-level approaches. The study also investigates zero-shot and few-shot prompting and ensemble majority voting. The experimental results show that LLaMA 3 achieves 86.7% accuracy with the two-stage sentence-level approach. The study demonstrates how different LLMs and prompting strategies can be evaluated for classifying non-hate, offensive, and hate speech in low-resource languages. It also provides a foundation for further research on multilingual hate speech detection.

REFERENCES

1. DIBYANAYAN BANDYOPADHYAY SAMRAT MUKHERJEE 1,BABANGAIN”M-NLE: Knowledge-Augmented Multitask Learning for Offensive Meme Detection and Explanation Generation”,11 September 2026.
2. HAIFA SALEH ALFURAYJ 1,2, DEWAN MD. FARID 3,(Senior Member, IEEE), CRISTINA LUNA-JIMÉNEZ”CYBY24 and Step-Wise Model for Thread-Based Fine-Grained Cyberbullying Detection”,22 January 2026.
3. EBTIHAL ENFES 1,NOOREDEEN AWWAD2, AND KOLAWOLE JOHN ADEBAYO”Cross-Dialectal Transfer Learning for Offensive Language Detection in Arabic”,April 2026.
4. NATHANAEL P. SOLAGRATIA ANDG.S.THEJAS”BERT and Local Vision Language Models in Hate Speech Detection on Small Multimodal Datasets”,May 2026.
5. AVITAL ROZENTHAL , AVIGAIL STEKEL,”A Systematic Analysis of Oversampling and Classic Machine Learning Pipelines for Multilingual Offensive Language Detection in Resource-Constrained Scenarios”,June 2026.
6. MAXIMILIAN NIENHAUS , MANUEL WEBER AND ANDREAS HUMPE,”Leveraging Conversational Context for German Hate Speech Detection With Large Language Models”,August 2026.
7. MAHMOUD ABUSAQER , (Graduate Student Member, IEEE), JAMIL SAQUER”BERT-OTA: Enhancing Hate Speech Detection With Ontology-Guided Transformer Attention”January 2026.
8. FARISHA AZMI1, NORMAISHARAH MAMAT SULTAN NOMAN QASEM 1,RAWADABDULGHAFOR”Culturally Aware Malay–English Code-Mixed Hate Speech Detection: A Systematic Review and Research Taxonomy”August 2026.
9. KATHIRAVAN PANNERSELVAM ANDSARANYARAJIAKODI,”IRAIVI: Agentic Framework for Tamil Hate Speech Detection and Counter-Speech Generation Toward Responsible AI Intervention”,September 2026.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details