JISTaP https://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 14(1): 87-100, 2026

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2026.14.1.6

Exploring the Role of Prompt Engineering and Large Language Models in Climate Change-Related Hate Speech Detection

Nishtha Kesswani
Department of Data Science and Analytics, Central University of Rajasthan, Ajmer, India
E-mail: nishtha@curaj.ac.in

Basant Agarwal
Department of Computer Science and Engineering, Central University of Rajasthan, Ajmer, India
E-mail: basant@curaj.ac.in

Krishna Kumar Mohbey*
Department of Computer Science, Central University of Rajasthan, Ajmer, India
E-mail: kmohbey@curaj.ac.in

Nikol Yunevich
Center of Advanced Studies in Digital Development, JSC "Giprosvyaz," Minsk, Belarus
E-mail: yunevich@giprosvjaz.by

Maksim Korneevets
Center of Advanced Studies in Digital Development, JSC "Giprosvyaz," Minsk, Belarus
E-mail: korneevets@giprosvjaz.by

ABSTRACT

Hate speech ought to be identified and minimized to diminish the undesirable influx of toxic content on the Internet, particularly in the fields of sensitive matters, such as the climate change debate. This article analyzes hate speech detection using climate change-related information and focuses on tweet classification and the identification of the hate target via prompt engineering with large language models (LLMs). Specifically, we compare three instruction-tuned LLMs, TinyLlama, Flan-T5, and Gemma, and evaluate them in contrast to a standard transformer baseline, bidirectional encoder representations from transformers (BERT). The aspects of the experimental framework include prompt design in a systematic way, reproducible implementation environments, and performance measures such as accuracy, precision, recall, and F1-score. The results suggest that prompt-based LLMs for zero-shot inference of hate speech do not require task-specific training. However, the finetuning BERT baseline is more precise and achieves a high F1-score, suggesting the importance of classification reliability. The findings indicate the pros and cons of prompt engineering, which is feasible but highlights the need for prompt design and model choice to achieve reliable hate speech detectors in climate-related speech.

Keywords: hate speech detection, social media, large language models, prompt engineering, climate change, natural language processing

Received: January 13, 2026 **Revised:** February 15, 2026
Accepted: February 25, 2026 **Published:** March 30, 2026

***Corresponding Author:** Krishna Kumar Mohbey
https://orcid.org/0000-0002-7566-0703
E-mail: kmohbey@curaj.ac.in

1. INTRODUCTION

Hate speech detection is a vital research area now, given its significant influence on society. The proliferation of social media such as Facebook and Twitter (now X) has changed the face of communication, but has also made it easier to spread harmful content. Nevertheless, such platforms are also gradually becoming a means of distribution of harmful and offensive content aimed at individuals or groups (Mohbey et al., 2025).

Hate speech is a term used to describe communication that offends, intimidates, or discriminates against a person or group of people based on their race, religion, gender, ethnicity, sexual orientation, or beliefs (Albladi et al., 2025; Mohbey et al., 2025). This kind of content may cause severe psychological, social, and emotional damage, as well as societal division and conflict. In addition, user anonymity makes the problems more difficult (Al-Maatouk et al., 2020). These actions can have a negative influence on the emotional and psychological condition of the abused. Hate speech detection has emerged as a significant research field because of its role in society (Jahan & Oussalah, 2023; Mossie & Wang, 2020). Social media has made communication and the exchange of information reach heights never seen before. Nevertheless, besides the benefits of virtual communication, hate speech has become more widely spread, which is extremely dangerous to personal and social integrity.

The conventional methods of hate speech identification mostly rely on keyword matching, lexical analysis, and conventional machine learning methods. Nevertheless, such methods do not frequently reflect context, sarcasm, implicit hate, or changing language patterns. However, recent deep learning technologies, in particular transformer models and large language models (LLMs), have demonstrated significant improvements in text comprehension (Mukherjee & Das, 2023). The instruction-tuned LLMs include TinyLlama, Flan-T5, and Gemma, which can reason contextually and categorize natural language when provided with structured prompts. Moreover, prompt engineering has proven to be a useful method for tasking LLMs, instructing them with structured prompts to enhance their performance in classification and interpretation.

The rise in sharing of climate change discourse on social media ecosystems has led to the escalation of polarization, the spread of false information, and the spread of goal-oriented hate speech towards individuals, activists, scientists, and communities. Climate change hate speech presents unique issues due to its domain-specific vocabulary, lack of contextual hints, and emotionally charged information. The existing automated hate speech detectors lack climate-specific optimization and thus face challenges with contextual interpretation. This is why there is an urgent need to explore new methodological solutions, including prompt engineering and LLMs, to improve the accuracy of detection and the interpretability of AI in this specific field. The primary objectives of this research are:

- To explore prompt engineering and instruction-tuned LLMs for hate speech detection in climate change–related social media tweets.
- To analyze the performance of various models (TinyLlama, Flan-T5, and Gemma) to identify their strengths and limitations.
- To highlight challenges in hate speech detection and provide directions for future research.

Although hate speech detection models, both traditional and transformer-based, have advanced, there are several gaps. The current methods are more dependent on supervised learning and are not usually useful in specific domains, such as climate change discourse. The systematic analysis of prompt engineering and its effects on classification and hate-target identification have not been studied in recent years, although recent studies have delved into the ellipses of LLMs. Moreover, there is little systematic research comparing prompt-guided LLMs to traditional transformer baselines. To fill these gaps, this paper assesses prompt-engineered LLMs for detecting climate change-related hate speech and compares their performance with a transformer-based model across a range of tasks.

To fill these gaps, the current paper will systematically test the application of prompt engineering methods on several LLMs, including TinyLlama, Flan-T5, and Gemma, and compares the quality of the results with those of a traditional transformer baseline (bidirectional encoder representations from transformers; BERT). In contrast to earlier research, this paper is dedicated to the problem of hate speech about climate change and analyzes the performance of models in a variety of tasks, such as hate speech recognition and hate target recognition. This study takes a step forward in creating adaptive, interpretive, and efficient hate speech detectors by offering empirical analysis of prompt-guided performance of LLMs in a domain-specific environment.

The rest of this paper is structured in the following way: Section 2 outlines the dataset and preprocessing

 https://doi.org/10.1633/JISTaP.2026.14.1.6

procedures. Section 3 is a detailed literature review. Section 4 explains the suggested methodology and prompt engineering strategy. Section 5 provides an analysis of the experiment's results. Lastly, Section 6 provides concluding remarks and suggests avenues for future research.

2. BACKGROUND AND PRELIMINARIES

The feature extractors common in traditional hate-speech detection methods are bag-of-words (feature extraction) (Zhang et al., 2010), term frequency-inverse document frequency (Sammut & Webb, 2011), and word-embedding-based feature extraction (Oro et al., 2025). In its turn, prompt-based LLMs are driven by an alternative paradigm. Rather than explicitly extracting features, pretrained language models implicitly encode linguistic and semantic and contextual evidence, which they learn by massive pretraining. In prompt-engineering engines, the model is directed through a natural-language interface, in the form of structured instructions, to do classification directly on raw text (Oro et al., 2025). This does not need manual feature engineering and makes it possible to do hate speech detection flexibly and at scale using pretrained contextual representations.

In the traditional model training for hate speech detection, feature extraction and data preprocessing are crucial steps performed on the labeled dataset before training or finetuning the model. These techniques involve transforming text data into numerical representations as features that can be fed into classification models. However, in prompt-based text generation approaches, such as those employed in prompt engineering techniques with LLMs, feature extraction and data preprocessing are not explicitly applied during model training or testing. Instead, the focus is on designing contextually relevant prompts that guide the model to generate hate-speech-related responses based on the given input prompt.

Contextual prompt design is a critical part of prompt engineering, as it shapes the model's responses and behavior. The prompts provide helpful information and clues that assist the LLMs in identifying patterns of hate speech and generating acceptable responses. Although feature extraction and data preprocessing are not directly applied to prompt-based text generation, the representations learned by LLMs and their internal processes contain linguistic features and contextual information, which makes it easier to produce responses that are meaningful and appropriate to the situation. Both feature extraction and data preparation are significant in the detection of hate speech. They are typically performed on labeled data during training or finetuning. Nonetheless, they are not applied in prompt-based text generation, including prompt engineering with LLMs. Rather, they are concerned with generating appropriate prompts and directing the model's answer generation process using the provided input prompt. This takes advantage of LLMs' natural ability to generate meaningful, relevant text in the context (Meguellati et al., 2025).

3. LITERATURE REVIEW

The problem of hate speech identification has become a rather important research topic with considerable social, psychological, and informational consequences of hate speech on the Internet space. Besides being a technical classification problem, hate speech is a socially constructed and contextual process that is determined by the specifics of language, platform-based norms, cultural definition, and the groups to which the hate speech is directed. Classification and information filtering are therefore carried through automated detection systems, which help in content moderation, trust, transparency, and decision-making. This review evaluates the traditional machine learning techniques, deep learning techniques, and emerging prompt-based LLM techniques critically and identifies key gaps that suggest the necessity of the present study.

3.1. Traditional and Deep Learning Approaches

The initial hate speech recognition systems used a rule-based approach, handcrafted linguistic attributes, and supervised machine learning models, including support vector machines, Naive Bayes, and maximum entropy classifiers. Other articles, such as Alsafari et al. (2020) and Al-Hassan and Al-Dossari (2022), created their own annotated datasets and used bag-of-words, lexicon-based representations, and feature-extraction techniques based on word embeddings. Such methods showed the possibility of automated hate speech classification and developed basic datasets and assessment systems. Nevertheless, conventional methods have a number of weaknesses. Their dependence on manual feature engineering limits their capacity to capture the contextual meaning, sarcasm, implicit hate, and changing linguistic patterns. Also, the models of this type usually consider hate speech as a lexical phenomenon as opposed to a dynamic and context-dependent informational phenomenon. These restrict how well they work in more intricate areas like climate change discussion, where there can be hate speech that is either indirect or entrenched in ideological messages.

Convolutional neural networks, long short-term memory networks, and transformer-based networks, in particular BERT and robustly optimized BERT pretraining approach, have demonstrated a great capability to enhance the performance of hate speech detection through deep learning. Vidgen et al. (2021) created large annotated datasets and proved that transformer models could effectively find fine-grained categories of hate when considering various social facets. In line with this, Mathew et al. (2021) have also pointed out the necessity to detect the target communities and offer explanations, which can be achieved only by having interpretable hate speech detectors.

Transformer models enhance context due to their ability to employ semantic relationships among words and phrases. Nevertheless, there are still a few drawbacks to these models. To begin with, they need a large amount of labeled training data, which may not be readily accessible in a domain-specific application such as climate change discourse. Second, the traditional transformer models are more like classification systems without necessarily having clear explanations and reasoning mechanisms. Third, they can have difficulty generalizing to new trends in hate speech due to static training processes. Such limitations point to the necessity of more adjustable, adaptive methods that are capable of accommodating changes in language patterns and can give interpretable outputs.

3.2. Domain-Specific and Climate-Related Hate Speech

Recent studies have generalized hate speech detection to domain-specific contexts, which include biomedical, scientific, and pandemic-related communication. Other researchers like Alshalan et al. (2020) have investigated hate speech during COVID-19 discourse, showing that conversations on the crisis might increase the use of harmful language. On the same note, there have been transformer-based models like biomedical BERT and scientific BERT that have been created to capture domain-specific language representations (Lee et al., 2019; Wei et al., 2019). These papers demonstrate that the task of hate speech detection should consider language peculiarities of the domain, contextual meaning, and social standards. Nevertheless, the bulk of the available literature focuses on general hate speech detection rather than more specific areas, such as climate change. The language of climate change has its own problems, such as polarization of ideology, indirect aggression against activists, and implicit use of harmful language. Moreover, there has been little research on the role of hate speech detection systems as information systems that facilitate transparency, auditability, and content moderation decisions. This void highlights the need for detection techniques that can not only categorize harmful content but also provide interpretable, actionable results.

The debate on climate change is now characterized by polarization, which leads to the formation of specific hate speech against the members of activist movements, policymakers, and social cohorts. EL-Sayed & Nasr (2024) and Kaya et al. (2024) established that transformer-based models are suitable in the detection of hate speech and stance in climate-related tweets. These papers have highlighted the relevance of contextual characteristics, including hashtags, mentions, and metadata on tweets. Nevertheless, previous work in climate-related hate speech detection is mainly based on traditional supervised transformer-based models. Such methods need a large amount of annotated data, and they cannot usually be adjusted to new hate speech forms. Also, they fail to critically analyse the effect of prompt design or model-interaction strategies on detecting performance. This weakness highlights the necessity of dynamic detection systems capable of managing changing climate-related discourse.

3.3. Prompt Engineering and Large Language Model-Based Detection

The recent progress in LLMs like GPT, LLaMA, and Flan-T5 has proposed novel hate speech detection paradigms featuring prompt engineering. Prompt-based models allow them to reason and classify according to natural-language instructions, in contrast to the more traditional supervised models. As Chain-of-Thought prompting is suggested to be effective in hate speech moderation, the HATE-GUARD model suggested by Vishwamitra et al. (2024) revealed the idea of chain of thought prompting. In the same way, Guo et al. (2023) tested the prompting strategies on benchmark datasets and demonstrated that prompt design has a significant impact on the performance of the LLM. The HARE proposed by Yang et al. (2023) is a model that uses the explanations generated by the LLM to enhance the accuracy and explainability of the detection. Although all these studies show the prospects of prompt-based approaches, a number of gaps are to be filled:

- Most studies do not provide systematic comparisons between prompt-guided LLMs and conventional transformer baselines.

 https://doi.org/10.1633/JISTaP.2026.14.1.6

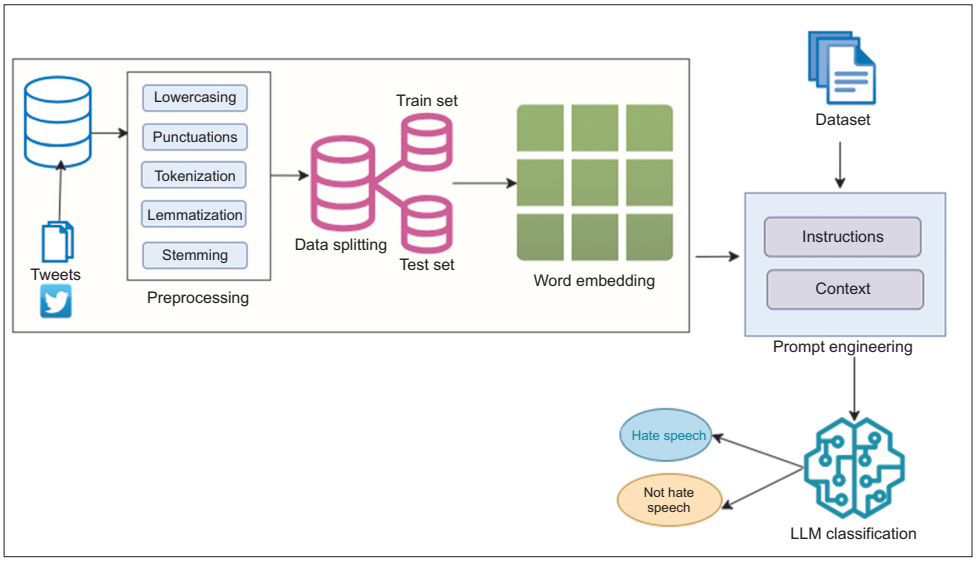

Fig. 1. Schematic framework of hate speech detection using prompt engineering and LLM. LLM, large language model.

- Limited research examines the effectiveness of prompt engineering in domain-specific contexts, such as climate change discourse.
- Few studies evaluate the ability of prompt-guided models to perform multiple tasks simultaneously, including hate speech classification and target identification.
- The relationship between prompt structure and model performance remains insufficiently analyzed.

These limitations highlight the need for empirical evaluation of prompt engineering strategies within domain-specific hate speech detection tasks.

4. METHODOLOGY

This paper explores the detection of hate speech and the identification of hate targets with prompt engineering on LLMs. Rather than finetuning models, a prompt-based classification model was used, where structured instruction prompts were used to prompt pretrained LLMs to classify tweets. The overall framework is illustrated in Fig. 1.

4.1. Dataset Description

The data utilized in this study were the ClimaConvo, as given by Shiwakoti et al. (2024). The data consists of Twitter posts about climate activism and environmental discussions obtained in social media. It marks hate speech, hate targets, and stance identification, and enables one to study harmful hate speech in its complexity in the setting of online climate discussions. In this paper, hate speech is speech that targets, threatens, or attacks groups or individuals due to certain characteristics that the group or individual is safeguarded or defended against in terms of characteristics such as race, religion, sex, or ideology. This definition is the same as that of dataset annotation and past studies on hate speech detection.

Table 1. Dataset composition

Category	No. of tweets
Total tweets	15,309
Relevant tweets	10,407
Non-relevant tweets	4,902

The hate speech detection and hate target identification experiments were conducted using the annotated subset of relevant tweets.

4.1.1. Dataset Structure

The ClimaConvo dataset consists of 15,309 tweets. Out of them, 10,407 tweets are marked as relevant to climate activism and environmental discourses, and 4,902 tweets are defined as non-relevant (Shiwakoti et al., 2024). The corresponding subset has hate speech detection, hate target detection, and stance detection annotations. Table 1 shows the entire dataset composition to make it transparent and traceable.

4.1.2. Dataset Split

Stratified sampling was used to separate the dataset into training, validation, and testing sets to maintain the same consistency of the classes in all the subsets. Namely,

70% of the data was assigned to the training set, 15% to the benchmark set, and 15% to the testing set. Even though the prompt-based LLMs were tested on the basis of zero-shot inference, the split of the data was preserved in order to be consistent with standard evaluation procedures and could be compared with supervised baseline models. There was no testing set that was applied to evaluate the final performance, except to have a legitimate and accurate assessment.

4.1.3. Tasks Definitions

The ClimaConvo data is favorable to various classification problems of harmful content detection. Subtask A is concerned with the detection of hate speech, whereby the tweets are classified as hate speech or non-hate speech using their content. Subtask B is the identification of the hate target, which is to determine the target or the group that is the subject of hate speech, like individuals, groups, or organizations. Subtask C will entail stance detection, where the stance as expressed in climate activism tweets is identified. The given research is mostly concerned with hate speech identification and hate target recognition.

4.2. Data Preprocessing

As used in this study, pretrained LLMs use prompts to classify data, which is why little preprocessing was done to ensure that the original semantic and contextual information of the tweets was preserved. Preprocessing was done to lowercase all the text and to remove formatting artifacts, and to keep useful contextual features, including hashtags, mentions, and punctuation that can carry useful semantic features (Aliero et al., 2023). The structured prompt template was planted with each tweet directly without the use of stopword removal, stemming, or feature engineering. This is to make sure that the models utilize what they have already trained in terms of contextual knowledge.

4.3. Model Selection and Baseline Configuration

This paper assesses prompt-based classification on instruction-tuned LLMs. The chosen models are TinyLlama-1.1B-Chat, Flan-T5-Small, and Gemma-2B-Instruct. The models were chosen because they exhibit instruction-following abilities and are well-suited to prompt-based natural language classification problems. Besides prompt-based models, a supervised BERT model was used as a baseline to identify hate targets. This baseline allows for comparing zero-shot classification via prompts with standard supervised learning methods. In contrast to the old-fashioned supervised techniques, the LLMs considered in this present paper were not trained on data. Rather, classification was based on structured prompt templates, and the models' classification capabilities were assessed.

4.4. Bidirectional Encoder Representations from Transformers Baseline Model

Hate target detection was done as a supervised baseline on the BERT model. BERT is a contextual representation-learning language model that is a pretrained transformer. The model was trained on labeled data and tested using standard classification metrics, including accuracy, precision, recall, and F1-score. This benchmark is used to compare the performance of prompt-based classification using LLMs (Naaz et al., 2021).

4.5. Prompt Engineering and Classification Framework

To tackle the problem of hate speech detection and hate target identification, LLMs were directed with the help of prompt engineering. The objective of classification, definitions of categories, and output constraints were well outlined through structured instruction prompts. All of the tweets were inputted into a prompt template, which was predefined and passed to the model. The model produced a textual answer with the projected classification label, which was directly retrieved as a part of the output and contrasted with the ground truth label. The methodology allows task-independent finetuning, and reproducible and consistent assessment between different models. Prompt-based classification enables the use of pretrained language models in a cost-effective way, and flexibility and interpretability in classification problems (Marvin et al., 2024).

5. EXPERIMENTS AND RESULTS

5.1. Experimental Setup

The experiments were made to measure the effectiveness of the prompt engineering method using LLMs in detecting hate speech and identifying hate targets. ClimaConvo data was stratified into training (70%), validation (15%), and testing (15%) sets. The test set was used to evaluate the prompt-based LLMs using zero-shot inference, and the dataset split was kept to be able to compare the results with the supervised baseline models. The test set was evaluated based on standard classification measures such as accuracy, precision, recall, and F1-score. Experimental conditions were the same, and structured prompt-based classification was used to evaluate all the

 https://doi.org/10.1633/JISTaP.2026.14.1.6

models. The model provided the classification decision without any further finetuning or classifiers. With this approach, there would be equal comparison of the models and uniform assessment of prompt engineering effectiveness across them.

5.2. Reproducibility and Implementation Details

To promote complete reproducibility and methodological transparency, detailed implementation instructions, such as immediate templates, preprocessing processes, hyperparameters, and execution environments, are given. The experiments were performed with pretrained LLMs, which are TinyLlama-1.1B-Chat-v1.0, Flan-T5-Small, and Gemma-2B-Instruct. They were implemented in HuggingFace Transformers and Keras-NLP, which are Python 3.10 frameworks. All the tweets were placed in an organized instructional prompt that was specially created to detect hate speech. The textual outputs produced by the models were texts with classification labels that were processed to retrieve the category that was predicted. The final evaluation did not rely on any external sentiment analysis or complementary classifiers. This is to make sure that the reported results are as per the inherent capabilities of the tested models and as per the efficacy of expedi-

Table 2. Complete prompt, hyperparameter, preprocessing, and execution configuration

Category	Parameter	Specification
Models	TinyLlama	TinyLlama-1.1B-Chat-v1.0
	Flan-T5	Google/flan-t5-small
	Gemma	Gemma-2b-instruct
Prompt configuration	Prompt type	Instruction-based zero-shot classification
	Task	Hate speech detection and hate target identification
	Output labels	Hate speech, non-hate speech
	Output constraint	Single label output
	Prompt format	Structured instruction template with explicit definitions and output requirements
Hyperparameters	Max_new_tokens	50
	Temperature	0.0
	Decoding strategy	Greedy decoding
	Num_return_sequences	1
Preprocessing	Text normalization	Lowercasing applied
	Special characters	Preserved to maintain contextual meaning
	Stopword removal	Not applied (to preserve semantic integrity)
	Lemmatization	Not applied (to preserve original linguistic structure)
	Dataset split	Stratified sampling: Train (70%), validation (15%), test (15%)
Execution environment	Programming language	Python 3.10
	Framework	HuggingFace Transformers, Keras-NLP, PyTorch
	Hardware	Graphics processing unit execution when available; otherwise central processing unit
	Tokenizer	Pretrained model-specific tokenizer
Inference settings	Input	Structured prompt+tweet text
	Output	Generated classification label extracted from model output
	External classifiers	Not used
Evaluation	Metrics	Accuracy, precision, recall, F1-score
	Evaluation method	Comparison with ground truth annotations using test dataset

ent engineering. The pretrained model-specific tokenizers were used to execute the experiments. Its implementation was done on GPU hardware (where it was available) or on CPU hardware. Each model was tested with the same prompt structures and decoding arrangements to ensure consistency and reproducibility. Deterministic decoding (temperature=0.0) was used on all experiments to have consistent and reproducible results.

Table 2 provides complete prompt specifications, preprocessing details, hyperparameters, and execution settings to ensure full reproducibility of the experiments. All experiments were conducted using fixed random seeds and deterministic decoding to ensure full reproducibility.

5.3. Model Selection for Hate Speech Detection

In the process of selecting the model to use in hate speech detection, we considered several factors to ensure that we settle on the best models to use in our objectives. After the deep analysis, we have chosen TinyLlama, Flan T5, and Gemma due to their unique advantages, compliance with timely engineering processes, and their ability to provide contextually suitable responses. This discussion gives a detailed discussion on the reasons why these models were selected:

5.3.1. TinyLlama

TinyLlama (Zhang et al., 2024) is a lightweight, resource-efficient framework designed for deployment in environments with limited resources, such as mobile or edge devices. TinyLlama is designed with prompt engineering in mind so that we can build contextually relevant prompts and can successfully direct the generation of answers by the model. TinyLlama is a strong contender for hate speech detection where efficiency is paramount, as it produces excellent, contextually pertinent responses even though the model is small.

5.3.2. Flan-T5

Flan-T5 is a state-of-the-art architecture that is characterized by its generality and efficiency when used in a variety of natural language processing applications. Flan T5 is built based on T5 (text-to-text transfer transformer) (Longpre et al., 2023). Prompts such as this give valuable indications to the model to identify hate speech, and we can make them personalized with the help of Flan T5, which is a prompt engineering model. Thanks to the scalability and flexibility of Flan T5, we are able to choose how to prompt model parameters and formulations to optimize the performance of our particular hate speech detection task.

5.3.3. Gemma

Gemma (Gemma Team, 2024) is an instruction-tuned LLM that is trained to follow natural language instructions and execute a broad spectrum of natural language processing tasks. Instruction tuning enhances the interpretation properties of the model to produce task-relevant responses to prompts. Gemma shows good contextual knowledge and is able to do classification tasks without task-specific finetuning based on prompt-based methods. It is instruction-following that makes it appropriate in assessing prompt-based hate speech detection.

5.4. Prompt Design and Configuration

Prompt engineering was the primary tool for steering LLMs toward hate speech detection. Under this method, tweets were placed in the context of a structured, instruction-based prompt that formally stated the classification task and the required output format. The task was given in a timely manner, with category names and clear output limitations that provided uniform and interpretable reactions of the model. In this work, a zero-shot prompt engineering technique was used. In zero-shot prompting, the model is not finetuned to a specific task during classification (Zhou et al., 2023). Instead, the pretrained model uses its existing linguistic and contextual information, along with the structured prompt, to produce classification decisions. In this way, it is possible to assess the intrinsic reasoning capacity of LLMs in detecting hate speech. The timely template was well laid out to make it very clear, reproducible, and consistent among all the models tested. The Twitter messages in the ClimaConvo dataset were given to the model as input by inserting the Twitter messages in the input prompt template. The models produced textual responses with classification labels, and they were then broken down to receive the category that was predicted.

The exact prompt template used in the experiments is presented in Fig. 2.

All of the test set tweets were, respectively, loaded into the {tweet} input, and the programmatically generated output was analyzed to obtain the prediction label of classification. This directive prompt makes sure that the task is well specified and the model generates a controlled and consistent output format. The clear definitions used in the prompt are useful in minimizing ambiguity and enhancing the reliability of classification. The experimental conditions were used to make sure that all of the models

Instruction:
You are an expert system for hate speech detection.
Task:
Determine whether the following tweet contains hate speech.
Definitions:
Hate Speech: Content that attacks, threatens, or targets individuals or groups based on protected characteristics such as race, gender, religion, or ideology.
Non-Hate Speech: Content that does not contain hateful, abusive, or targeted harmful language.
Tweet:
"{tweet}"
Output Requirement:
Return only one label:
Hate Speech
or
Non-Hate Speech

Fig. 2. Prompt template for hate speech detection. Structured prompt template used to guide the model for hate speech classification of tweets during the experiments.

were consistent and could be compared to each other, and the identical prompt template, format structure, and output limits were applied to all of them, such as TinyLlama, Flan-T5, and Gemma. No prompt-related specifics of the models were added. The model-generated output was used to make the classification decision without referencing external classifiers or other supporting sentiment analysis tools. This guarantees that the assessment of the intrinsic ability of the LLMs and prompt engineering performance has been represented.

5.5. Evaluation Metrics

The model's performance was evaluated based on precision, recall, F1-score, and accuracy (Mohbey et al., 2025), as shown in equations (1)-(4).

$$Precision = \frac{True_{pos}}{True_{pos}+False_{pos}} \quad (1)$$

$$Recall = \frac{True_{pos}}{True_{pos}+False_{Neg}} \quad (2)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (3)$$

$$Accuracy = \frac{True_{pos}+True_{Neg}}{True_{pos}+True_{Neg}+False_{pos}+False_{Neg}} \quad (4)$$

5.6. Results

The present section provides experimental findings on the utilization of prompt-based classification on LLMs to detect hate speech (Subtask A) and hate targets (Subtask B). The classification decisions were simply obtained as the results of the model outputs using the structured prompt template as discussed in Section 5.4. The ground truth annotations were compared with their predicted labels to calculate accuracy, precision, recall, and F1-score. Notably, the final assessment did not utilize any external classifier, sentiment-based analysis models, or any other scoring mechanisms. This guarantees that the reported results are based on the classification ability of the tested LLMs in the specified prompt setting and the performance of prompt engineering.

5.6.1. Hate Speech Detection Results

5.6.1.1. TinyLlama Model.

The TinyLlama-1.1B-Chat model was tested on the structured instruction-based prompt template as outlined in Section 5.4. The prompt included each of the tweets, and the model produced a text reply with the estimated classification label. The ground-truth annotation was then compared with the predicted label, which was directly derived from the model output, and the evaluation metrics were computed. The TinyLlama model had 80% accuracy, 25% precision, 50% recall, and 33% F1-score. These findings imply that TinyLlama has a strong overall classification accuracy and moderate recall, implying that it can be used to detect hate speech cases. Nevertheless, the comparatively low accuracy implies the existence of false positives, which also points to the necessity of additional advances by timely refining the models or moving to larger and instruction-tuned ones.

5.6.1.2. Flan-T5 Model.

The Flan-T5-Small model was tested in the same structure of instruction-based prompt template and the same experimental setup so that the models are fairly and equally compared. The prompt included each of the tweets, and the model produced a text reply with the estimated classification label. The generated output was specially extracted into the predicted label that was compared with the ground truth annotation to calculate the evaluation metrics. The accuracy, precision, recall, and F1-score of the Flan-T5 model were 80%, 25%, 33% and 29% respectively. These findings show that Flan-T5 has competitive overall classification accuracy, which is similar

to TinyLlama. Nevertheless, it has a lower recall and F1-score, indicating lesser accuracy in the consistent detection of hate speech cases. This difference in performance can be explained by the smaller size of the model and the sensitivity of the model to quick interpretation, which can influence the reliability of classification in more complex language situations (Table 3).

5.6.2. Hate Target Detection Results

The task in Subtask B was to recognize the hate speech target of tweets. To determine the effectiveness of both traditional transformer-based models and prompt-based LLMs in hate speech target detection, both models were tested. The analysis was conducted through the comparison of the obtained predicted labels to the ground truth annotations in terms of traditional performance measures, such as accuracy, precision, recall, and F1-score.

5.6.2.1. Bidirectional Encoder Representations from Transformers Model.

The BERT-based sequence classification model recorded good and uniform performance in all the metrics of evaluation. The model had an accuracy, precision, recall, and F1-score of 0.68, 0.69, 0.68, and 0.68, respectively. These findings reveal that the BERT model is able to give dependable, balanced classification performance on hate-target detection. The accuracy is very high, implying that the model will generate relatively low false positive predictions; as well, the model has a high recall, which means that it is useful in accurately predicting hate speech victims. The overall effectiveness and robustness of the model are also proven by the balanced F1-score.

5.6.2.2. Gemma Model.

Gemma-2B-Instruct model was tested with the same structured instruction-based prompt template as given in Section 5.4. All of the tweets were incorporated into the prompt, and the model produced a textual response with the predicted classification label. The ground-truth annotation was directly compared with the predicted label, which was directly recovered from the model output. The Gemma model had an accuracy of 0.68, precision of 0.53, recall of 0.54, and F1-score of 0.53. These findings suggest that Gemma performs on average with comparatively even-handed accuracy and recall. Although its performance is worse than the finetuned BERT model, Gemma shows that instruction-tuned LLMs can be used to perform classification tasks with prompt-based methods without task-specific finetuning.

5.6.2.3. TinyLlama and Flan-T5 Models.

These models were also evaluated for hate-target detection using the same prompt-based classification framework. Since the distribution of predictions was similar in zero-shot settings and they exhibited prompt-response behavior, their evaluation measures were similar to those in Subtask A. This consistency indicates similar response patterns of the model and not evaluation artifacts (Table 4).

5.7. Comparative Analysis and Discussion

The experimental results reveal a distinct performance trade-off between traditional supervised learning and zero-shot prompt-based inference. To better contextualize these findings within the existing literature, Table 5 provides a comparative performance matrix that summarizes how our results align with or diverge from prior studies.

Table 3. Results for Subtask A (hate speech detection)

Model	Accuracy	Precision	Recall	F1-score
TinyLlama	0.80	0.25	0.50	0.333
Flan-T5	0.80	0.25	0.333	0.286

Table 4. Results for Subtask B (hate target detection)

Model	Accuracy	Precision	Recall	F1-score
BERT	0.68	0.69	0.68	0.68
TinyLlama	0.80	0.25	0.50	0.333
Flan-T5	0.80	0.25	0.333	0.286
Gemma	0.68	0.53	0.54	0.53

BERT, bidirectional encoder representations from transformers.

 https://doi.org/10.1633/JISTaP.2026.14.1.6

Table 5. Comparative performance matrix of evaluated models

Feature	BERT (supervised baseline)	Prompt-guided LLMs (TinyLlama/Flan-T5/Gemma)
Data dependency	High: Requires large, domain-specific labeled datasets for finetuning	Minimal: Relies on pre-trained weights; operates via zero-shot/few-shot prompts
Classification precision	Superior (~0.69): Lower false-positive rate; captures nuances of hate targets	Limited (~0.25): High false-positive rate due to conservative "safety" overgeneralization
Domain adaptability	Static: Requires retraining for new domains (e.g., from generic to climate hate)	Dynamic: Adaptable via rapid prompt adjustments and instructional wording
Interpretability	Opaque: Provides labels but lacks inherent reasoning or textual justification	High: Capable of generating "step-by-step" reasoning (chain-of-thought) for labels
Resource demand	High compute for training; low compute for inference	Low setup overhead; high compute for inference (parameter-heavy)

BERT, bidirectional encoder representations from transformers; LLM, large language model.

The outcomes of the experiments present the key performance indicators of the prompt-based LLMs in the hate speech detection task. Although TinyLlama and Flan-T5 had comparatively high accuracy (0.80), they produced low precision, which means that they had false positives. This implies that prompt-based models can overgeneralize in detecting hate speech, especially in a zero-shot scenario. The patterns of prediction and the distribution of classes that exhibit the dominant one can be pointed out as the reason for the same values of accuracy that all the models indicate. Nevertheless, precision, recall, and F1-score offer a more detailed measure of the classification performance. The reduced precision values in prompt-based models point to the importance of better prompt design and model calibration. The finetuned BERT model, on the contrary, had more balanced performance with higher values of precision and F1-score. This shows that finetuning can help the model to gain more accurate task-specific representations and enhance classification accuracy.

The Gemma model showed moderate results, which validated the hypothesis that prompt-based methods on instruction-tuned LLMs are capable of classification. Nevertheless, they are not as good as finetuned models, indicating the trade-off between prompt-based flexibility and supervised finetuning. These results indicate that prompt engineering offers an effective and scalable method of hate speech detection, but it is also necessary to pay attention to the design of prompts and careful model choice to obtain high-quality performance. Attributes of data sets and distribution patterns of classes affect the same accuracy values realized in some prompt-based models. Accuracy does not necessarily give the complete picture of classification reliability, especially in imbalanced classification environments. Precision, recall, and F1-score differences also allow a better understanding of model behavior and emphasize the differences in the capacity to label the instances of hate speech as correct. The differences in accuracy and precision observed imply the presence of class imbalance and conservative predictive behavior, which may lead to high overall accuracy and, in turn, false-positive predictions.

Our findings confirm the observations of Guo et al. (2023) and Vishwamitra et al. (2024), who noted that while LLMs possess vast linguistic knowledge, their zero-shot performance in "adversarial" or "niche" domains (such as climate change) often lags specialized, finetuned models. The high accuracy (0.80) paired with low precision (0.25) in TinyLlama and Flan-T5 suggests a majority-class dominance effect—a phenomenon often reported in imbalanced social media datasets where models default to "non-hate" labels to maximize accuracy while failing to recall the minority "hate" class effectively.

6. ETHICAL CONSIDERATIONS

When choosing and adopting models, ethical considerations should be of the highest priority. Solutions were established to guarantee fairness, transparency, and privacy in hate speech detection, such as countermeasures to reduce bias and ethical principles in the use of data and consent. This integrated model selection and implementation plan provides a systematic approach to the implementation of timely engineering methods using LLMs to detect hate speech, the need to structure relevant contextual prompts, transparent and reproducible evaluation processes, and ethical concerns during the investigation

process. Using this methodology, we will aim to develop a powerful hate-speech detection model that can assist in creating safer and more welcoming online spaces.

7. CONCLUSION

This paper examined the performance of prompt engineering methods used on the LLMs, including TinyLlama, Flan-T5, and Gemma, on hate speech detection and hate target identification problems. The findings of the experiment prove that prompt-based classification can allow LLMs to detect hate speech without any task-specific finetuning. The tested models proved that prompt-based models are feasible in resource-constrained environments to detect hate speech, but their accuracy and F1 scores were less than those of the finetuned BERT baseline. The results show that instruction-based prompts, which are organized, are of paramount importance in modeling behavior and facilitating characterization. Although conventional finetuned transformer models like BERT had better overall performance, prompt-based LLMs models were able to effectively conduct classification tasks by using zero-shot inference. This emphasizes the versatility and usefulness of prompt engineering, especially in situations where training data or computation components are scarce and labeled. Nevertheless, the findings also suggest imperfections of accuracy and classification reliability of smaller instruction-tuned models, showing possibilities of improvement by designing better prompts, larger model structures, and better instruction tuning. Further studies can seek to identify strategies to optimize the prompt in real-time and few-shot prompting methods, as well as test models on larger, more heterogeneous datasets to enhance the future strength and generality of models. On the whole, this paper shows that timely engineering is a useful, reproducible mechanism of using LLMs in hate speech recognition, which helps to design more efficient and flexible content moderation systems.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGEMENTS

This work was supported by the Department of Science and Technology India, International Cooperation Division (Grant No: DST/INT/BLR/P-38/2023(G)), and the Belarusian Republican Foundation for Fundamental Research (Agreement No. F23INDG-011).

REFERENCES

Al-Hassan, A., & Al-Dossari, H. (2022). Detection of hate speech in Arabic tweets using deep learning. *Multimedia Systems*, 28(6), 1963-1974. https://doi.org/10.1007/s00530-020-00742-w

Al-Maatouk, Q., Othman, M. S., Aldraiweesh, A., Alturki, U., Al-Rahmi, W. M., & Aljeraiwi, A. A. (2020). Task-technology fit and technology acceptance model application to structure and evaluate the adoption of social media in academia. *IEEE Access*, 8, 78427-78440. https://doi.org/10.1109/ACCESS.2020.2990420

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*, 13, 20871-20892. https://doi.org/10.1109/ACCESS.2025.3532397

Aliero, A. A., Adebayo, B. S., Aliyu, H. O., Tafida, A. G., Kangiwa, B. U., & Dankolo, N. M. (2023). Systematic review on text normalization techniques and its approach to non-standard words. *International Journal of Computer Applications*, 185(33), 44-55. https://doi.org/10.5120/ijca2023923106

Alsafari, S., Sadaoui, S., & Mouhoub, M. (2020). Hate and offensive speech detection on Arabic social media. *Online Social Networks and Media*, 19, 100096. https://doi.org/j.osnem.2020.100096

Alshalan, R., Al-Khalifa, H., Alsaeed, D., Al-Baity, H., & Alshalan, S. (2020). Detection of hate speech in COVID-19–related Tweets in the Arab region: Deep learning and topic modeling approach. *Journal of Medical Internet Research*, 22(12), e22609. https://doi.org/10.2196/22609

El-Sayed, A., & Nasr, O. (2024, March 22). AAST-NLP at ClimateActivism 2024: Ensemble-based climate activism stance and hate speech detection : Leveraging pretrained language models. In A. Hürriyetoğlu, H. Tanev, S. Thapa, & G. Uludoğan (Eds.), *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)* (pp. 105-110). Association for Computational Linguistics.

Gemma Team; Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., … Eck, D. (2024). *Gemma: Open models based on Gemini research and technology*. https://arxiv.org/abs/2403.08295

Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., & Hu, H. (2023, December 15-17). An investigation of large language models for real-world hate speech detection. In M. Arif Wani, M. Boicu, M. Sayed-Mouchaweh, P. H. Abreu, & J. Gama (Eds.), *2023 International Conference on Machine Learning and Applications (ICMLA)* (pp. 1568-1573). IEEE.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Kaya, A., Ozcelik, O., & Toraman, C. (2024, March 22). ARC-NLP at ClimateActivism 2024: Stance and Hate speech detection by generative and encoder models optimized with tweet-specific elements. In A. Hürriyetoğlu, H. Tanev, S. Thapa, & G. Uludoğan (Eds.), *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)* (pp. 111-117). Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., & Roberts, A. (2023, July 23-29). The flan collection: Designing data and methods for effective instruction tuning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 22631-22648). JMLR.

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In I. Jeena Jacob, S. Piramuthu, & P. Falkowski-Gilsk (Eds.), *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2023* (pp. 387-402). Springer Singapore.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A benchmark dataset for explainable hate speech detection. *IAAI-21, EAAI-21, AAAI-21 Special Programs and Special Track*, 35(17), 14867-14875. https://doi.org/10.1609/aaai.v35i17.17745

Meguellati, E., Pratama, N., Sadiq, S., & Demartini, G. (2025, April 28-May 2). Are large language models good data preprocessors? In G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, H. Huang, & E. Yom-Tov (Eds.), *Companion Proceedings of the ACM on Web Conference 2025* (pp. 2129-2132). Association for Computing Machinery.

Mohbey, K. K., Agarwal, B., Kesswani, N., Sterjanov, M., Nikol, Y., & Margarita, V. (2025). Hate speech identification and categorization on social media using Bi-LSTM: An information science perspective. *Journal of Information Science Theory And Practice*, 13(1), 51-69. https://doi.org/10.1633/JISTaP.2025.13.1.4

Mossie, Z., & Wang, J. H. (2020). Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3), 102087. https://doi.org/10.1016/j.ipm.2019.102087

Mukherjee, S., & Das, S. (2023). Application of transformer-based language models to detect hate speech in social media. *Journal of Computational and Cognitive Engineering*, 2(4), 278-286. https://doi.org/10.47852/bonviewJCCE2022010102

Naaz, S., Abedin, Z. U., & Rizvi, D. R. (2021). Sequence classification of tweets with transfer learning via BERT in the field of disaster management. *EAI Endorsed Transactions Scalable Information Systems,* 8(31), e8. https://doi.org/10.4108/eai.23-3-2021.169071

Oro, E., Granata, F. M., & Ruffolo, M. (2025). A comprehensive evaluation of embedding models and LLMs for IR and QA across English and Italian. *Big Data and Cognitive Computing*, 9(5), 141. https://doi.org/10.3390/bdcc9050141

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Shiwakoti, S., Thapa, S., Rauniyar, K., Shah, A., Bhandari, A., & Naseem, U. (2024, May 20-25). Analyzing the dynamics of climate change discourse on Twitter: A new annotated corpus and multi-aspect classification. In N. Calzolari, M. Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 984-994). ELRA and ICCL.

Vidgen, B., Thrush, T., Waseem, Z., & Kiela, D. (2021, August 1-6). Learning from the worst: Dynamically generated datasets to improve online hate detection. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1667-1682). Association for Computational Linguistics.

Vishwamitra, N., Guo, K., Romit, F. T., Ondracek, I., Cheng, L., Zhao, Z., & Hu, H. (2024, May 19-23). Moderating new waves of online hate with chain-of-thought reasoning in large language models. *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 788-806). IEEE.

Wei, C. H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, 47(W1), W587-W593. https://doi.org/10.1093/nar/gkz389

Yang, Y., Kim, J., Kim, Y., Ho, N., Thorne, J., & Yun, S. Y. (2023, December 6-10). HARE: Explainable hate speech detection with step-by-step reasoning. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5490-5505). Association for Computational Linguistics.

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). *TinyLlama: An open-source small language model*. https://arxiv.org/abs/2401.02385

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1), 43-52. https://doi.org/10.1007/s13042-010-0001-0

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). *Large language models are human-level prompt engineers*. https://arxiv.org/abs/2211.01910