PLD at HOMO-LAT 2025: Enhancing Dialectal Sentiment Analysis through Contextual Retrieval and Translation

Jose Olivert-Iserte[†], Diego Caballero-García-Alcaide[*,†] and Lucía Guijarro-Martínez

Computer Science and Engineering Department, Carlos III University of Madrid, Spain

Abstract

In the digital age, online social media platforms have transformed human communication, but they have also become venues for the propagation of hate speech, particularly targeting marginalized communities such as the LGBTQ+ population. In this context, Natural Language Processing (NLP) offers valuable tools for the automatic detection and classification of harmful content. The HOMO-LAT25 shared task, part of the IberLEF 2025 evaluation campaign, focuses on sentiment analysis toward the LGBTQ+ community in Latin American Spanish. It includes two subtasks: a multi-dialect track with consistent training and evaluation dialects, and a cross-dialect track, which challenges systems to generalize sentiment classification across unseen dialects. This study addresses the linguistic and cultural complexities inherent in dialectal variation, which significantly affect how sentiment is expressed and perceived. To tackle these challenges, several Transformer-based approaches using pre-trained models such as BERT and its domain-adapted variants are proposed. Additionally, some innovative enhancements are introduced, including context-based techniques and Retrieval-Augmented Generation (RAG), to incorporate semantically relevant information during inference. These methods aim to improve robustness and accuracy in polarity detection, ultimately contributing to safer and more inclusive online spaces for LGBTQ+ communities.

Keywords

Hate Speech Detection, NLP, Text Classification, RAG, Context-Based Inference, LLM, BERT, Prompting

1. Introduction

In the digital era, online social media has revolutionized how we communicate, interact, and share ideas. Unfortunately, this transformation has also led to harmful applications of online spaces, such as the propagation of hate speech. More specifically, the LGTBQ+ community frequently experiences such aggression, facing constant targeting on these online platforms [1].

In this context, Natural Language Processing (NLP) offers a powerful tool for the automatic detection of online hate speech [2]. Shared tasks play a crucial role in this field by providing common datasets, evaluation metrics, and a collaborative environment to foster advancements. The Iberian Languages Evaluation Forum (IberLEF) [3] is a shared evaluation campaign for NLP systems which focuses on Spanish and Iberian languages. Its aim is to promote the research of text processing, understanding, and generation, fostering innovation and advancing the state-of-the-art in NLP for these language communities through the organization of competitive shared tasks.

Within the framework of its 2025 edition, the HOMO-LAT25 [4] shared task is dedicated to studying societal perceptions of the LGBTQ+ community through language on social media, as its predecessors [5, 6], focusing on polarity detection (sentiment analysis) in Latin American Spanish. It requires systems to classify posts as having positive, negative, or neutral sentiment towards the LGBTQ+ community or related keywords across two distinct settings: a multi-dialect track (Subtask 1), where training and evaluation dialects are consistent, and a cross-dialect track (Subtask 2), which introduces the complexity of differing dialects between training and evaluation, testing model generalization. A primary objective, central to the work, is to address the complexities caused by such linguistic diversity. These dialectal

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

✉ jolivert@inf.uc3m.es (J. Olivert-Iserte); dicaball@inf.uc3m.es (D. Caballero-García-Alcaide); 100455351@alumnos.uc3m.es (L. Guijarro-Martínez)

iD 0009-0000-9031-5644 (J. Olivert-Iserte); 0009-0005-4451-9642 (D. Caballero-García-Alcaide)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-10

variations significantly influence how sentiment, particularly negative and potentially phobic language, is expressed and perceived, posing a challenge for developing systems sensitive to local slang and culturally specific expressions. Thus, the core challenge is to accurately identify targeted sentiment robustly across diverse linguistic forms, especially in the cross-dialect setting.

Successfully addressing this challenge through the development of models capable of robust and precise polarity detection is intended to foster a safer online environment for Latin American LGBTQ+ individuals.

To address these challenges, this work proposes several approaches based on Transformer architectures, specifically leveraging pre-trained models such as BERT and its domain-adapted variants for social media. In addition, innovative techniques are explored to enhance model predictions, including Retrieval-Augmented Generation (RAG) and context-based strategies that incorporate semantically relevant information during inference to improve robustness and accuracy, particularly in the presence of dialectal variation. These approaches and enhancements not only demonstrated strong empirical performance but also led to the system achieving the top position in the HOMO-LAT25 shared task, validating the effectiveness of the proposed methodologies.

2. State of the Art

2.1. Transformers

Transformers [7] have marked a revolutionary advancement within the field of NLP and other areas of Artificial Intelligence. This architectural paradigm has proven to be exceptionally effective for tasks demanding sequence comprehension, including machine translation, text generation, classification, and sentiment analysis. In contrast to prior models such as Recurrent Neural Networks (RNNs) [8], Transformers employ an attention mechanism that captures long-range dependencies more efficiently, enabling them to outperform earlier approaches across a diverse array of tasks.

The fundamental architecture of Transformers primarily consists of two main components: an encoder and a decoder. The encoder processes an input sequence, transforming it into a rich feature representation. Subsequently, the decoder uses this representation to generate an output sequence. Both the encoder and decoder blocks are constructed from layers incorporating attention mechanisms and fully connected feed-forward neural networks [9]. Central to this design is the attention mechanism [10]. It empowers the model to selectively focus on different segments of the input sequence by enabling each element to relate to others through calculated attention weights. These weights determine the relative importance of each part of the sequence, thus enhancing the model's capacity to capture long-range dependencies and consider the global context of the input.

This architecture has been indispensable to virtually all significant advancements in AI in recent years. Despite their success, Transformer-based models also come with limitations [11]. Firstly, they demand substantial computational resources for both training and inference. Secondly, the complexity of these models often impedes their interpretability, introducing challenges for applications where transparency is a critical requirement.

2.2. LLMs

Large Language Models (LLMs) have emerged as one of the most significant innovations in the field of Artificial Intelligence in recent years. These models are fundamentally built upon Transformer-based architectures, designed to capture complex, long-range relationships within data, as previously mentioned [12]. Furthermore, LLMs are trained on vast text corpora, through which they acquire an implicit understanding of linguistic patterns, context, and semantic relationships.

The capabilities of Large Language Models have significantly impacted various NLP tasks, sentiment analysis being one among them [13]. Traditional sentiment analysis methods often relied on feature engineering, lexicons, or simpler machine learning models, which could struggle with the complexities of language, such as emotions, attitudes, and implicit opinions [14].

LLMs, with their deep understanding of language acquired from pre-training on massive datasets, offer several advantages for sentiment analysis. They can often perform sentiment classification with high accuracy even in zero-shot or few-shot settings, meaning they can determine sentiment without requiring task-specific training data, or with very little [15]. This is particularly beneficial for domains or languages where labeled sentiment data is scarce.

However, the use of LLMs in sentiment analysis also presents disadvantages. In the context of this work, like their general use, they can inherit biases present in their training data, potentially leading to skewed sentiment predictions for certain demographic groups or topics [16]. Nonetheless, the integration of LLMs has significantly advanced the capabilities and accuracy of sentiment analysis systems, enabling a more nuanced and context-aware understanding of opinions expressed in text.

2.3. Aspect-Based Sentiment Analysis

Beyond determining the overall polarity of a text, Aspect-Based Sentiment Analysis (ABSA) provides a more fine-grained analytical approach [17]. The core objective of ABSA is to identify specific aspects, defined as attributes or components of an entity or topic mentioned within the text, and subsequently ascertain the sentiment expressed towards each of these distinct aspects [18].

This process typically involves deconstructing the text into its constituent pieces, identifying the core elements or aspects under examination, and subsequently analyzing the sentiment expressed towards each of these [18, 19]. The capacity of LLMs to understand complex linguistic structures and contextual relationships has notably enhanced the performance of ABSA systems, particularly in accurately identifying subtle aspect mentions and their associated sentiment. Thus, ABSA offers a detailed and structured understanding of opinions, moving beyond general sentiment scores. In the context of tasks like HOMO-LAT25, this approach could further identify the specific themes or topics within pertinent discussions that drive particular sentiments, offering deeper insights into the characteristics of problematic or supportive language.

2.4. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has become a key strategy to enhance the relevance, adaptability, and factual grounding of LLM outputs across diverse domains [20, 21]. In sentiment analysis, particularly for nuanced or domain-specific contexts such as social media discourse, RAG helps address limitations in the model's internal knowledge by retrieving semantically relevant examples or documents from curated repositories before generation. This retrieved context is then incorporated into the input prompt, grounding the model's predictions in up-to-date and task-specific data [22].

RAG systems typically consist of a knowledge base, embedding models for retrieval, vector databases (e.g., FAISS, Pinecone), and a retrieval pipeline [23, 22]. In the case of sentiment classification, such systems can retrieve similar past posts labeled with sentiment classes, user feedback, or even related community guidelines, providing the model with more informed context. Fine-tuning embeddings on sentiment-labeled data and applying prompt engineering strategies further improve retrieval accuracy and classification performance.

Recent innovations, such as hybrid keyword-semantic retrieval and multimodal RAG approaches, enable more context-aware outputs, particularly useful in emotionally complex or dialectally diverse settings [24, 21]. Moreover, integrating feedback mechanisms allows the system to iteratively refine its retrieval and classification pipeline, improving both consistency and robustness [25].

2.5. Related Work

Multiple studies have addressed the challenge of identifying various forms of online hate, with part of this work specifically focusing on content targeting the LGBTQ+ community. Early approaches often relied on lexicon-based methods and traditional machine learning classifiers [26]. However, recent research has increasingly leveraged deep learning, particularly Transformer-based models, to better capture contextual nuances.

For instance, [27] presented work on detecting homophobia and transphobia in YouTube comments across English and Tamil, introducing a new dataset and benchmarking various machine learning models. Their study, highlighted the challenges and successes in recognizing such content across different linguistic contexts. Similarly, [28] focused on identifying LGBT+-directed hate speech in tweets using multi-class and multi-label approaches with models like BERT and RoBERTa. They specifically addressed practical challenges such as data imbalance through preprocessing and oversampling techniques. Further exploring Transformer capabilities, [29] employed a GPT-2 model for recognizing homophobic and transphobic content in social media comments across five languages, including Spanish, noting variance in detection ease across languages. Addressing the complexities of multilingual and low-resource scenarios, [30] investigated the detection of homophobia and transphobia in code-mixed social media text. They proposed data augmentation techniques, such as pseudo labeling through transliteration, to improve model performance, demonstrating strategies to tackle resource scarcity in this domain. These studies collectively underscore the shift towards sophisticated neural models and the ongoing efforts to tackle the nuanced and evolving nature of online homophobia, often with a multilingual perspective.

For processing Spanish-language social media text, robust pre-trained language models are crucial. RoBERTuito [31] has emerged as a significant resource in this regard. Developed by pre-training a RoBERTa architecture on a large corpus of Spanish tweets (over 500 million), RoBERTuito is specifically tailored to understand the informal language, slang, and idiosyncrasies prevalent in social media conversations in Spanish. Its strong performance on various downstream tasks, including sentiment analysis and offensive language identification in Spanish, has made it a widely adopted baseline and a foundational model for researchers working with Spanish social media data, often serving as a starting point for fine-tuning in tasks related to sentiment and hate speech detection.

3. Datasets

This section provides a detailed description of the data made available by the organizers for the tasks introduced in Section 1. A total of four datasets were provided for training and evaluating the systems developed for the competition.

Table 1
Training and Development datasets number of instances and features.

Dataset	Instances	Features
Training	5767	5
Development	1443	5

The datasets summarized in Table 1 correspond to the training and development partitions used for Track 1, which focuses on multi-dialect sentiment analysis as outlined in Section 1. These datasets were released at separate times and contain user-generated content labeled with sentiment and dialect metadata. Each instance includes five features: *id*, *country*, *keyword*, *post content* and *label*. Specifically, four dialects are represented in both sets: Argentinean, Colombian, Chilean, and Mexican Spanish. The dialects remain fixed in Track 1, which allows models to specialize in these regional variations.

The challenge escalates in complexity when transitioning to Track 2, the cross-dialect setting, where the distribution of dialects in the test data includes variations not present in the training corpus. This divergence introduces a domain adaptation issue, further complicating model generalization.

Figure 1 illustrates the class distribution across the training and development sets. As shown, there is a marked imbalance, with the positive (POS) class being significantly underrepresented. This imbalance poses well-known difficulties for machine learning models, such as a tendency to overfit on the majority classes and poor generalization for minority ones, which directly impacts performance in real-world applications.

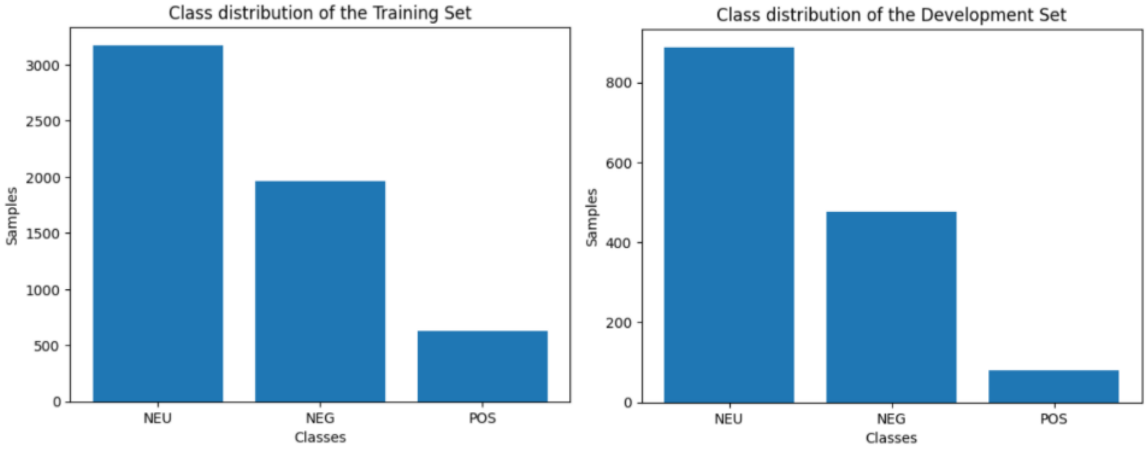

Figure 1: Class distribution for Training and Development sets.

Figure 2 shows the distribution of text lengths across the dataset. As shown, in both the training and development sets, 95% of the texts contain fewer than 300 words, with only a small number of instances exceeding this length. This observation is particularly relevant for determining an appropriate maximum token limit that the models can efficiently handle during training and inference.

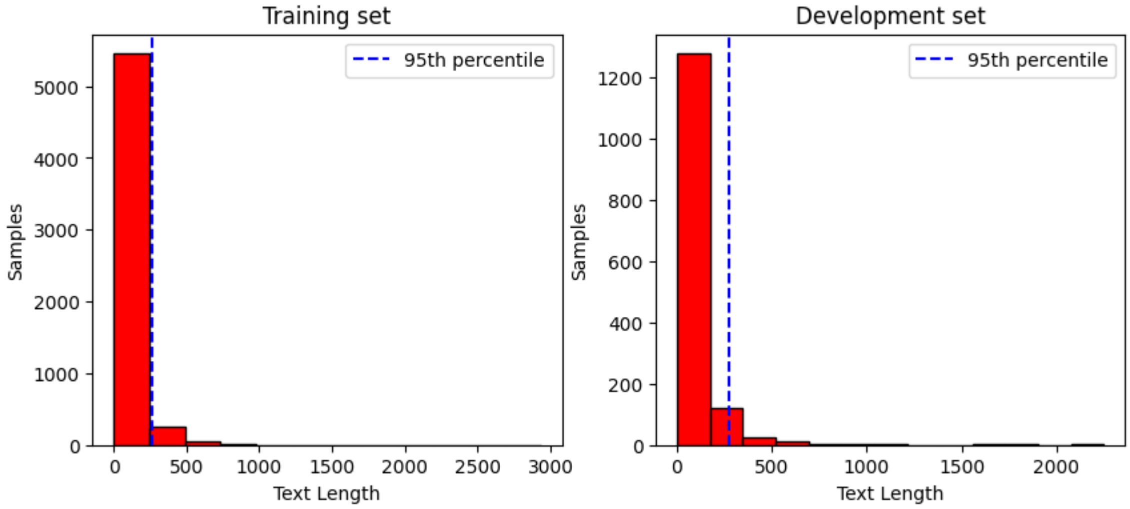

Figure 2: Texts length for Training and Development sets.

In addition to the training and development sets, two separate test sets were provided for evaluation. The first test set, corresponding to Track 1 (cross-dialect), comprises 3,588 instances. The second test set, associated with Track 2 (multi-dialect), contains 5,475 instances.

It is worth noting that the number of instances in the test sets is approximately equal to the total number of training and development instances. This parity in size ensures a balanced evaluation scenario; however, it also necessitates that models exhibit strong generalization capabilities, as the evaluation phase involves datasets of comparable complexity and volume to those used during training.

4. Methodology

This section outlines the methodology adopted to tackle the tracks presented in the HOMO-LAT25 challenge. It begins by detailing the preprocessing procedures applied to the original dataset, followed by the data augmentation strategies implemented. Subsequently, the context extraction is described and some conclusions are presented with an overview of the various approaches explored to address the problem.

4.1. Data Processing and Augmentation

To begin with, as the texts originated from Reddit posts, it is performed a cleaning of each entry in the original dataset. This preprocessing step involved the removal of superfluous and redundant elements such as URLs, user mentions, markdown syntax, and extra whitespaces. Additionally, all text was converted to lowercase in order to minimize vocabulary size.

Subsequently, all texts were translated into English. This decision was motivated by two key considerations:

- **Pre-trained models:** The search for open-source pre-trained models revealed that the majority, particularly those demonstrating superior performance, were trained predominantly on English-language data. Even multilingual models were largely biased toward English, with Spanish comprising only a minor portion of the training data.
- **Spanish dialectal variation:** Given that the second track includes Spanish dialects in the test set that are not represented in the training data, translating all texts into English prior to model training and inference served as an effective strategy to mitigate dialectal discrepancies.

Therefore, the hypothesis is that leveraging high-quality English pre-trained models, combined with the mitigation of dialectal variation through translation, would significantly enhance model performance across both tracks, particularly in Track 2.

Following the preprocessing stage, online data augmentation is implemented due to the limited number of training examples. Firstly, each text was paraphrased twice in sequence, with the goal of ensuring that the second paraphrase differed as much as possible from both the original and the first paraphrased version. Note that all paraphrased texts were also generated in English, based on the previously translated data.

During training, each input instance was randomly selected from a set comprising the original text and its two paraphrased versions. This strategy ensured that the model was exposed to slightly varied data in each epoch, thereby enhancing generalization and reducing the risk of overfitting.

It is worth noting that both the translation of the texts and the generation of paraphrases were performed using a Large Language Model, specifically GPT-4o-mini.

4.2. Context Engine

Most of the methodologies presented in the following section incorporate a novel component referred to as the Context Engine. As the name implies, this module is designed to retrieve semantically relevant contextual information to support the prediction of a given post. Due to the inherent complexity and subjectivity of the task at hand, as well as challenges such as class imbalance and limited data availability, providing the model with contextual examples prior to inference has proven to be a robust and effective strategy.

The approaches that leverage the Context Engine will be detailed in the subsequent section. In general terms, the component retrieves the top-k most semantically similar posts from the training dataset for each test instance. For example, when $k = 1$, the methodology retrieves the most semantically similar post from each class (NEG, NEU, and POS) with respect to the post being predicted. As will be demonstrated, this strategy not only enhances model performance but also highlights the importance of incorporating relevant and class-specific context in tasks that are highly subjective. To this end, cosine similarity is employed as the similarity metric. Cosine similarity is widely adopted in natural language processing tasks for quantifying the degree of similarity between high-dimensional vectors [32].

In this setting, similarity computations are not performed directly on raw text, but rather on embedding vectors, dense numerical representations of the input text. These embeddings are generated using the SentenceTransformers library [33], which provides state-of-the-art pre-trained models optimized for various semantic similarity tasks, including semantic search and paraphrase identification.

To improve efficiency and reduce computational overhead, the embeddings corresponding to all posts in the training corpus are precomputed and stored locally. The same embedding model is applied consistently to both the training data and the posts being evaluated during inference, ensuring semantic alignment and vector compatibility. When a new post is presented for classification, it is encoded into a 384-dimensional embedding. Cosine similarity is then computed between this embedding and all precomputed training embeddings, and the top-k most similar posts from each class are retrieved. These selected examples are subsequently used to provide context during the inference process.

To illustrate the functionality of the Context Engine with a real-world example, consider the following post to be classified:

> `“being gay and moving to the middle east is not a very bright idea much less contacting someone through an app knowing that it is illegal i do not justify their medieval customs but you dont have to be very bright to realize that they handed themselves on a silver platter to the authorities.”`

When applying the methodology with $k = 1$, the Context Engine retrieves one semantically similar post from each class. For instance, from the POS class, the retrieved post is: *“nowadays yes the country is promoted as a gay friendly destination attracting tourists from that community and dollars come in.”* From the NEG class: *“i wouldnt use tinder even if i were straight in the middle east now imagine being gay there.”* And from the NEU class: *“I don’t understand why homosexuals consciously go to these places that hate them. What is worse, it makes them illegal just for being homosexual. I hope no gay person approaches the Middle East from now on.”* These contextual examples are then used in different ways in order to add contextual information to the different approaches and models used.

4.3. Approaches

This section presents a comprehensive analysis of the approaches evaluated during the HOMO-LAT25 challenge. In total, four distinct methodologies are described. As an initial step, a baseline model is introduced to establish a performance reference using the preprocessed dataset. This model serves to demonstrate the fundamental capabilities of a standard architecture when applied to the task.

Subsequently, three alternative approaches are proposed, each of which outperforms the baseline in terms of predictive performance. These methods incorporate advanced techniques from the field of NLP and are further enhanced through the integration of novel strategies, including the use of contextual information and ensemble mechanisms. The specific design and implementation details of these enhancements are discussed in the following subsections.

4.3.1. Base Line Models

To establish an initial performance benchmark, two well-established machine learning algorithms were implemented as baseline models: Random Forest (RF) and Support Vector Machine (SVM). These models were selected due to their proven effectiveness in classification tasks, their relative interpretability, and their ability to provide competitive performance without the need for extensive hyperparameter tuning or high computational overhead. Their robustness and simplicity make them suitable for exploratory analyses and for serving as reference points in comparative evaluations.

The task at hand involves the detection of homophobic comments in Reddit posts, specifically within the Latin American sociolinguistic context. This introduces several challenges related to language variation, informal syntax, code-switching, and the presence of culturally nuanced expressions. While these characteristics complicate the task for traditional models, they also underscore the value of establishing a strong baseline to assess the complexity of the dataset before applying more advanced neural or context-aware architectures.

In this setting, both RF and SVM were trained on vectorized textual inputs derived from embedding representations of the Reddit posts. These embeddings capture semantic and syntactic features of the text, facilitating the application of classical classifiers that rely on fixed-length input vectors. Although these models do not inherently exploit sequential information beyond the word-level features captured by the embeddings, they offer an useful perspective on the discriminative power of the raw data.

Preliminary evaluation using these models helped to gauge the inherent difficulty of the classification task. It also revealed some of the limitations of traditional approaches when applied to imbalanced and noisy datasets, as is often the case with user-generated content in online platforms. Specifically, the class imbalance between homophobic and non-homophobic comments adversely affected the performance of both classifiers, resulting in high accuracy but suboptimal recall for the minority class, which is a critical metric in hate speech detection tasks. These initial findings served as a foundation for the development and evaluation of more advanced models discussed in subsequent sections.

4.3.2. LLM Prompting

The following approach leverages the GPT-4-o-mini model to perform sentiment classification based on a tailored natural language prompt. This stage directly follows the execution of the Context Engine (See Section 4.2), which is responsible for enriching the input instance with class-specific contextual information retrieved from the training corpus. The integration of contextual examples into the prompt is key to addressing the subjectivity of the classification task and improving model generalization, particularly in imbalanced or data-scarce environments.

The constructed prompt follows a standardized template designed to align the model's response with the sentiment classification task. Specifically, the prompt frames the model as an assistant whose sole responsibility is to determine the polarity, Positive (POS), Negative (NEG), or Neutral (NEU), of a Reddit post in relation to a given keyword. The complete format of the prompt is shown below:

```
You are an assistant tasked with classifying Reddit posts based on the
sentiment they express towards a given Keyword. There are three possible
classes: Positive (POS), Negative (NEG), or Neutral (NEU). Classify this
Reddit post as POS, NEG, or NEU based on its sentiment or polarity towards
this given keyword. Keyword: [keyword] Post to classify: [target post]
Return only the label. To give you more context, I will provide three
lists containing the most similar texts to the post to be classified.
Examples of positive posts: [list of k POS posts] Examples of negative
posts: [list of k NEG posts] Examples of neutral posts: [list of k NEU
posts] Remember, return only the label: POS, NEG, or NEU.
```

This approach ensures that the model is exposed not only to the raw post but also to semantically aligned examples from each sentiment class. These examples are the output of the Context Engine, as explained in Section 4.2.

Once the tailored prompt is constructed, it is passed to the GPT-4-o-mini model for inference. The model then outputs a single label, POS, NEG, or NEU, based on its interpretation of the post's sentiment with respect to the provided keyword and contextual examples. As illustrated in Figure 3, this output constitutes the final classification result of the entire system.

By embedding semantic context into the prompt, this approach significantly enhances the model's ability to handle subtle sentiment variations, particularly in complex or ambiguous posts. As will be demonstrated in the evaluation section, this methodology consistently outperforms baseline models, underscoring the critical role of contextual prompting in sentiment classification tasks involving social media data.

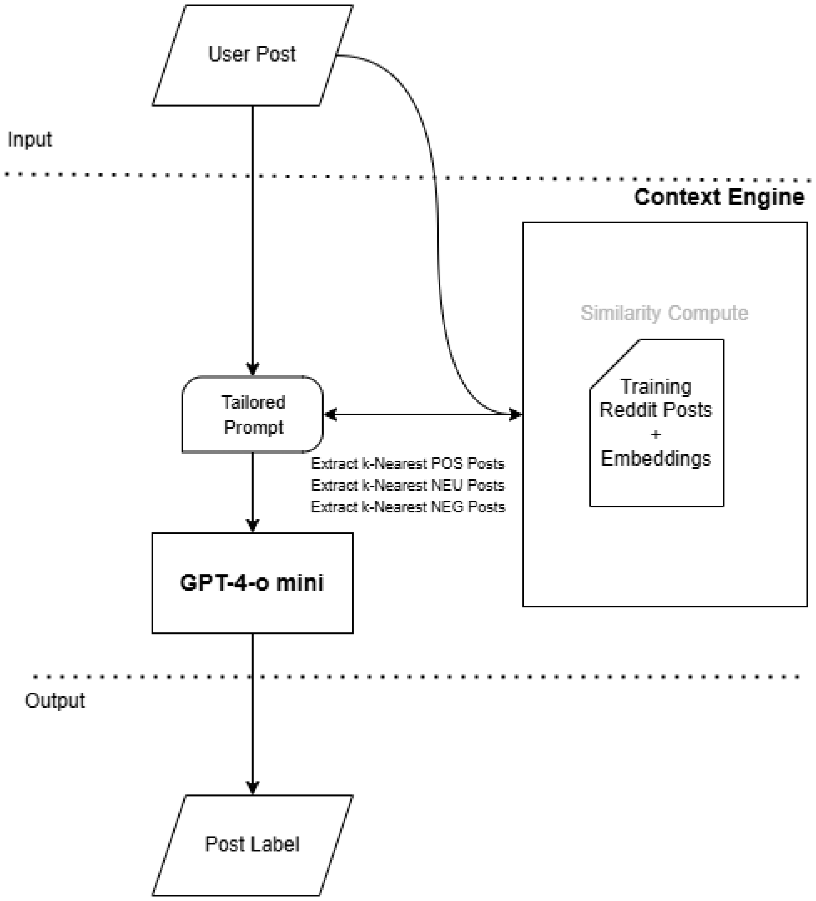

Figure 3: LLM Prompting Overview.

4.3.3. BERT Finetuning

This approach involves fine-tuning a BERT-based model. Specifically, an open-source text classification model available on Hugging Face: "cardiffnlp/twitter-roberta-base-sentiment-latest" [34] is selected. This is a RoBERTa-base model trained on 124 million tweets from January 2018 to December 2021, and finetuned for sentiment analysis. Although some experimentation was made with other models, this one consistently yielded superior results, likely due to the linguistic and stylistic similarities between tweets and Reddit posts.

Given that the task at hand is aspect-based sentiment analysis, a modification to the input structure was made to ensure the model focuses on the relevant keyword within each post. During tokenization, each input instance is constructed by concatenating the text and the corresponding keyword, separated by a designated separator token, as follows:

```
['<s>', TEXT-TOKENS, '</s>', '</s>', KEYWORD-TOKENS, '</s>']
```

This structure encourages the model to attend more closely to the text tokens that overlap with the keyword tokens during prediction.

The pre-trained model is fine-tuned using the training data, systematically exploring various hyperparameter configurations and training strategies to identify the setup that delivered the best performance. To further enhance the robustness and generalization of the model, an ensemble technique has been applied to this approach. As described in Section 4.1, the strategy involves generating paraphrases of each input post and performing inference not only on the original text but also on its two paraphrased variants. Each of these three versions is independently processed through the model, yielding three predicted sentiment labels. The final classification decision is then determined by majority voting among these predictions. In instances where the three predictions yield distinct labels, the final prediction is assigned to the label with the highest associated probability.

This ensemble method helps to mitigate the impact of potential ambiguities or artifacts present in a single formulation of the post, effectively smoothing out variability introduced by lexical or syntactic differences. As will be shown in the evaluation section, this strategy leads to improved performance and greater stability compared to relying solely on the original post for inference.

4.3.4. Context-Based BERT

This approach also leverages the Context Engine. However, in contrast to the LLM Prompting strategy, which injects these examples directly into a natural language prompt, this method integrates the contextual information into the architecture of a fine-tuned transformer-based classifier through a specialized attention mechanism. Specifically, it employs the "cardiffnlp/twitter-roberta-base-sentiment-latest" model, mentioned in the BERT finetuning approach, as the backbone and enhances it with class-specific multi-head attention modules.

The proposed model, referred to as "SentimentClassifierWithMultiAttention", extends a pre-trained RoBERTa encoder with three parallel multi-head attention blocks, each responsible for integrating contextual signals from a different sentiment class: Positive (POS), Neutral (NEU), and Negative (NEG). These blocks are crucial for tailoring the representation of the input post based on its semantic alignment with prototypical examples from each sentiment class.

During inference, the following sequence of operations is executed:

1. The input post is encoded using the RoBERTa backbone, and the output corresponding to the `[CLS]` token is extracted as the global representation of the post, denoted as $\mathbf{h}_{\text{cls}} \in \mathbb{R}^d$.
2. Using the Context Engine, k nearest posts from each class are retrieved. These posts are also encoded via RoBERTa to obtain their respective `[CLS]` embeddings:
   - $\mathbf{H}_{\text{pos}} \in \mathbb{R}^{k \times d}$
   - $\mathbf{H}_{\text{neu}} \in \mathbb{R}^{k \times d}$
   - $\mathbf{H}_{\text{neg}} \in \mathbb{R}^{k \times d}$
3. For each class $c \in \{\text{pos}, \text{neu}, \text{neg}\}$, a multi-head attention module is applied where:

$$\text{Attention}_c(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$$

   Here, $\mathbf{Q} = \mathbf{h}_{\text{cls}}$, $\mathbf{K} = \mathbf{H}_c$, and $\mathbf{V} = \mathbf{h}_{\text{cls}}$, making the attention mechanism focus on how much the representation of the input post is aligned with the embeddings of the respective class. This design ensures that the attention weights capture how the semantics of the post correlate with prototypical sentiment examples, but the value vector remains grounded in the input post's own representation.
4. The outputs of the three attention modules $\mathbf{a}_{\text{pos}}$, $\mathbf{a}_{\text{neu}}$, and $\mathbf{a}_{\text{neg}}$ represent contextualized versions of $\mathbf{h}_{\text{cls}}$, modulated by each sentiment class.
5. These four vectors are concatenated to form a unified representation:

$$\mathbf{h}_{\text{combined}} = \left[\mathbf{h}_{\text{cls}}; \mathbf{a}_{\text{pos}}; \mathbf{a}_{\text{neg}}; \mathbf{a}_{\text{neu}}\right] \in \mathbb{R}^{4d}$$

6. This combined vector is then passed through a deep feedforward network with dropout and non-linear activations. Finally, a linear classifier projects it into a 3-dimensional output corresponding to the sentiment classes.

This architecture allows the model to not only consider the content of the input post but also to explicitly attend to semantically similar posts from each class. By modeling interactions between the input and class-specific exemplars via attention, the system gains a structured inductive bias that enhances robustness, particularly under the presence of ambiguous or borderline cases. As with the LLM Prompting approach, the selection of contextual posts is powered by cosine similarity over SentenceTransformer embeddings. This shared mechanism ensures consistency in the semantic grounding across architectures while allowing each model to exploit context in a way best suited to its inference paradigm. In addition, to further enhance the robustness and generalization of the model, the same ensemble technique applied in BERT Finetuning approach (Section 4.3.3) has been applied to this approach.

5. Evaluation

This section presents the evaluation methodology applied to all the approaches described in Section 4.3, along with a discussion of the corresponding results. It is important to highlight that the evaluation procedure for the baseline models differs from that employed for the subsequent approaches. Specifically, a train-test split was conducted using the training dataset provided to the participants at the outset of the competition. Also, an exhaustive grid search was conducted to explore all possible hyperparameter configurations, with the objective of identifying the optimal settings for each model in terms of macro F1-score performance. This preliminary evaluation of the baseline models served to establish a reference point and to assess the initial difficulty of the task. The final reported results correspond to the official evaluation conducted during the HOMO-LAT25 competition, encompassing both Track 1 and Track 2.

Despite their general applicability, both baseline models exhibited limited effectiveness across both tracks, primarily due to the significant class imbalance present in the dataset. As shown in Tables 2 and 3, both Random Forest and SVM struggled to correctly classify instances from the minority class (POS), with an F1-score of 0.00. This result is indicative of the models' bias toward the majority classes, particularly the NEU class, which dominates the dataset.

Table 2
Classification report of the Random Forest (RF) model.

Class	Precision	Recall	F1-Score	Support
NEG	0.63	0.18	0.27	392
NEU	0.58	0.95	0.72	634
POS	0.00	0.00	0.00	127
Accuracy			0.58	1153
Macro avg	0.40	0.37	0.33	1153
Weighted avg	0.53	0.58	0.49	1153

Table 3
Classification report of the Support Vector Machine (SVM) model.

Class	Precision	Recall	F1-Score	Support
NEG	0.54	0.42	0.47	392
NEU	0.62	0.82	0.70	634
POS	0.00	0.00	0.00	127
Accuracy			0.59	1153
Macro avg	0.38	0.41	0.39	1153
Weighted avg	0.52	0.59	0.55	1153

The Random Forest model achieved an overall accuracy of 58%, with a macro-average F1-score of 0.33. The SVM model slightly improved the overall accuracy to 59% and the macro-average F1-score to 0.39. However, the zero recall and precision for the POS class in both cases suggest that these models are not suitable for handling highly imbalanced and subjective tasks without further enhancements such as resampling strategies, class weighting, or the integration of contextual information.

These results confirm the limitations of conventional machine learning algorithms in scenarios characterized by strong class imbalance and semantic complexity. The inability of these models to detect any instance of the POS class highlights the need for more advanced techniques that can incorporate external contextual knowledge and exploit semantic similarities between posts. As will be shown in subsequent sections, the proposed Context Engine addresses these shortcomings by enriching the input space with relevant and class-balanced contextual examples, significantly improving performance across all classes.

From now on the results obtained on the test set provided by the organizers of HOMO-LAT25 are presented and analyzed. In order to obtain robust and generalizable results, it is essential to explore how different hyperparameter configurations affect model performance. Testing a wide range of settings allows us to better understand the sensitivity of the models to various training dynamics and to identify the most optimal combinations for this specific task. This is particularly relevant in a challenging context such as sentiment classification of social media posts in the Latin American space, where linguistic variability and data imbalance introduce additional complexity.

Each approach was trained and evaluated under different combinations of these parameters, allowing us to conduct a thorough analysis of their impact on model effectiveness. The results of these experiments and the final configuration are discussed in the following subsections. In Table 4 the different parameters that have been tested in the different approaches can be seen. For context, *max_len* refers to the maximum number of tokens that the model can get as input, *d_aug* parameter refers to a binary variable to control whether to use data augmentation techniques, as explained in Section 4.1, or not, and parameter *k* refers to the number of posts of each class extracted by the Context Engine.

Table 4
Parameters tested for each approach.

Parameter	BERT Finetuning	LLM	Context-Based BERT
max_len	{128, 256, 512}	-	{128, 256, 512}
batch_size	{8, 16, 32}	-	{8, 16, 32}
epochs	{10, 15, 20}	-	{10, 15, 20}
learning_rate	{1e-6, 1e-5, 1e-4, 2e-5}	-	{1e-6, 1e-5, 1e-4, 2e-5}
weight_decay	{0.01, 0.1}	-	{0.01, 0.1}
d_aug	{0, 1}	-	{0, 1}
k	-	{2, 10}	{2, 10}

A total of five different submissions were made for the task, and the results for both tracks are summarized in Table 5. Each submission corresponds to a distinct approach, which is described as follows:

- **BERT**: Fine-tuned BERT model with ensemble strategy, trained on 90% of the dataset and validated on the remaining 10% to select the best-performing model during training. The final configuration of this approach is: *max_len* = 256, *batch_size* = 16, *epochs* = 10, *learning_rate* = 2e-5, *weight_decay* = 0.01 and *d_aug* = 1.
- **BERT-FULL**: Same as the BERT approach, but trained on the full dataset. The selected model corresponds to the epoch that yielded the best results in validation during prior tuning. The final configuration is the same as in **BERT** approach.
- **LLM**: Predictions generated using a prompting-based LLM approach.
- **CONTEXT**: Context-enhanced BERT model, trained on 90% of the data and evaluated on the remaining 10%. The final configuration of this approach is: *max_len* = 256, *batch_size* = 16, *epochs* = 10, *learning_rate* = 1e-5, *weight_decay* = 0.01, *d_aug* = 1 and *k* = 2.
- **CONTEXT-ENSEMBLE**: Context-enhanced BERT model with ensemble strategy, trained on 90% of the data and validated on the remaining 10%. The final configuration is the same as in **CONTEXT** approach.

Table 5
Test set results sorted by macro F1-score in descending order for Track 2.

Approach	Track 1	Track 2
BERT-FULL	0.5260	**0.5086**
BERT	0.5284	0.5051
LLM	**0.5296**	0.4989
CONTEXT-ENSEMBLE	0.5085	0.4961
CONTEXT	0.4946	0.4870

For Track 1, the best-performing approach was the LLM prompting method, achieving a macro F1-score of approximately 0.53. Both BERT-based fine-tuning approaches produced slightly lower yet comparable scores. In contrast, the context-based models obtained the lowest results, with F1-scores close to 0.50. In Track 2, the highest macro F1-score was achieved by the BERT-FULL approach, reaching approximately 0.51. Interestingly, although LLM prompting was the best approach in Track 1, it performed worse than both BERT-based methods in this track. Once again, the context-based models recorded the lowest performance, although the gap was smaller compared to Track 1.

A few additional observations can be made from the results:

- The BERT-FULL model, which was trained on the full dataset, outperformed the BERT model trained on a reduced dataset in Track 2. However, the opposite occurred in Track 1, though the difference in scores is minimal and likely not statistically significant.
- The use of ensemble strategies consistently led to improved performance across both BERT and context-based models, highlighting the benefit of aggregating predictions over relying on a single instance.

Additionally, Table 6 presents the results obtained by all participants in the HOMO-LAT25 challenge for both tracks. The data used to generate this table were sourced from the official results released by the organizers, which are available at the following link.

Table 6
All results sorted by macro F1-score in descending order for Track 1 (Left) and Track 2 (Right).

Owner	Track 1 (F1-Score)	Owner	Track 2 (F1-Score)
Ours (LLM)	0.5296	**Ours** (BERT-FULL)	0.5086
Ours (BERT)	0.5284	**Ours** (BERT)	0.5051
user1	0.5261	**Ours** (LLM)	0.4989
user1	0.5261	**Ours** (CONTEXT-ENSEMBLE)	0.4961
Ours (BERT-FULL)	0.5260	**Ours** (CONTEXT)	0.4870
user2	0.5137	user1	0.4803
user2	0.5137	user1	0.4803
user2	0.5104	user2	0.4639
Ours (CONTEXT-ENSEMBLE)	0.5085	user2	0.4639
Ours (CONTEXT)	0.4946	user2	0.4639
user1	0.4820	user1	0.4467
user3	0.4661	user4	0.4388
user4	0.4360	user5	0.4054
user5	0.4301	user5	0.4054
user5	0.4301	user3	0.3622
user3	0.3793	user6	0.2588
user6	0.2592		

The most noteworthy observation from Table 6 is the substantial performance gap observed in Track 2, where all five of the submitted approaches outperformed those of the other participants. In particular, the top-performing method surpassed the next best result by nearly 0.03 in macro F1-score.

This performance advantage is likely attributable to the strategy of translating all texts into English. Given that Track 2 features Spanish dialects in the test set that differ from those in the training data, a decline in performance was expected due to the increased linguistic variability. However, by translating the texts into English, the dialectal variation and linguistic noise is effectively reduced. This translation step contributed to greater textual homogeneity, thereby improving the model's ability to generalize across dialects and ultimately enhancing performance in this more challenging track.

6. Conclusions and Future Work

This paper presented several Transformer-based approaches to tackle the HOMO-LAT25 shared task, focusing on sentiment analysis towards the LGBTQ+ community in Latin American Spanish, with a particular emphasis on handling dialectal variations.The core strategies involved translating input texts to English to leverage powerful pre-trained models and mitigate dialectal discrepancies, employing a novel Context Engine to inject semantically relevant examples during inference, and utilizing data augmentation and ensemble techniques to enhance robustness.

The findings demonstrate the efficacy of these methods. Notably, the strategy of translating all texts to English proved highly successful for the cross-dialect track (Track 2), where all five of the submitted models outperformed other participants, with the top-performing BERT-FULL model achieving a significant lead. This underscores the benefit of standardizing linguistic input when dealing with diverse, low-resource dialects. For the multi-dialect track (Track 1), LLM prompting enriched with the Context Engine yielded the best results among the submissions, highlighting the power of large models when provided with carefully curated contextual information. Fine-tuned BERT models also showed strong, competitive performance across both tracks. The Context Engine, by providing class-specific similar examples, consistently aided models in navigating the subjective nature of sentiment classification.

These results not only highlight the promise of Transformer-based models and LLMs in tackling complex sentiment analysis tasks in low-resource and linguistically diverse settings, but also reflect a broader insight: while LLMs are powerful and increasingly indispensable tools in NLP, their effectiveness depends heavily on how they are integrated into task-specific pipelines. Relying solely on their pre-trained knowledge can lead to suboptimal outcomes, especially in domains where context, cultural nuance, or dialectal variation plays a central role. As such, future systems should not only adopt LLMs but also incorporate strategic components, such as contextual retrieval mechanisms, translation pipelines, or task-aware augmentation, to fully unlock their potential. A thoughtful preliminary analysis remains essential to tailor these models to the unique demands of each task.

For future work, several avenues warrant exploration. First, while English translation was effective, investigating multilingual models specifically pre-trained or fine-tuned on diverse Spanish dialects could offer a more direct approach, potentially preserving nuances lost in translation. Second, the Context Engine could be enhanced by exploring more sophisticated retrieval mechanisms beyond cosine similarity or by dynamically adjusting the number of retrieved examples (k) based on input characteristics. Integrating true Retrieval-Augmented Generation with external knowledge bases containing dialect-specific slang or cultural context could further improve performance. Third, a deeper analysis into model interpretability is crucial, especially for socially sensitive tasks like hate speech detection, to understand how models arrive at their predictions and to identify potential biases. Finally, it is worth noting that the dataset provided for the shared task was limited in size. Expanding the training data, both in quantity and dialectal coverage, would be highly beneficial, as it would allow for more robust fine-tuning and better generalization across diverse linguistic scenarios.

Acknowledgments

This work has been partially carried out with the support of grant PID2022-140554OB-C32, funded by MCIN/AEI/10.13039/501100011033, and project SEGVAUTO 5Gen-CM (TEC-2024/ECO-277), funded by the Community of Madrid. Additionally, this project has been partially supported by the strategic research initiative of the Carlos III University of Madrid, titled “Digital Engineering and Language AI”.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 and Google Gemini in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

[1] O. Ştefăniţă, D. M. Buf, Hate speech in social media and its effects on the lgbt community: A review of the current research1, 2021. doi:10.21018/rjcpr.2021.1.322.

[2] S. Agarwal, A. Sureka, Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website (2017).

[3] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[4] G. Bel-Enguix, H. Gómez-Adorno, S. Ojeda-Trueba, G. Sierra, J. Barco, E. Lee, J. Dunstan, R. Manrique, Overview of HOMO-LAT at IberLEF 2025: Human-centric polarity detection in Online Messages Oriented to the Latin American-speaking lgbtq+ populaTion, Procesamiento del lenguaje natural 75 (2025) –.

[5] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of HOMO-MEX at Iberlef 2023: Hate speech detection in Online Messages directed Towards the MEXican spanish speaking LGBTQ+ population, Procesamiento del lenguaje natural 71 (2023) 361–370.

[6] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, J. Andersen, Scott Thomas Vásquez, , S. Ojeda-Trueba, T. Alcántara, M. Soto, M. Cesar, Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection Towards the Mexican Spanish speaking LGBT+ Population, Procesamiento del lenguaje natural 73 (2024) 393–405.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 2017-December, Neural information processing systems foundation, 2017, pp. 5999–6009.

[8] S. Grossberg, Recurrent neural networks, Scholarpedia 8 (2013) 1888. doi:10.4249/scholarpedia.1888.

[9] G. Bebis, M. Georgiopoulos, Why network size is so important, IEEE Potentials 13 (1994) 27–31. doi:10.1109/45.329294.

[10] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning (2025). doi:10.1016/j.neucom.2021.03.091.

[11] C. Sanford, D. J. Hsu, M. Telgarsky, Representational strengths and limitations of transformers, 2023.

[12] H. Naveed, A. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. Mian, A comprehensive overview of large language models (2023). doi:10.48550/arXiv.2307.06435.

[13] W. Zhang, Y. Deng, B.-Q. Liu, S. J. Pan, L. Bing, Sentiment analysis in the era of large language models: A reality check, in: NAACL-HLT, 2023. URL: https://api.semanticscholar.org/CorpusID:258866189.

[14] B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, 2015. doi:10.1017/CBO9781139084789.

[15] S. Fatemi, Y. Hu, A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis, ArXiv abs/2312.08725 (2023). URL: https://api.semanticscholar.org/CorpusID:266210291.

[16] S. Poria, D. Hazarika, N. Majumder, R. Mihalcea, Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research, IEEE Transactions on Affective Computing 14 (2023) 108–132. doi:10.1109/TAFFC.2020.3038167.

[17] C. Wu, B. Ma, Z. Zhang, N. Deng, Y. He, Y. Xue, Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models, ArXiv abs/2412.12564 (2024). URL: https://api.semanticscholar.org/CorpusID:274789129.

[18] M. Wankhade, C. Kulkarni, A. C. S. Rao, A survey on aspect base sentiment analysis methods and challenges, Applied Soft Computing 167 (2024) 112249. URL: https://www.sciencedirect.com/science/article/pii/S1568494624010238. doi:https://doi.org/10.1016/j.asoc.2024.

112249.

[19] I. A. Kandhro, F. Ali, M. Uddin, A. Kehar, S. Manickam, Exploring aspect-based sentiment analysis: an in-depth review of current methods and prospects for advancement, Knowledge and Information Systems 66 (2024) 3639–3669. URL: https://doi.org/10.1007/s10115-024-02104-8.

[20] M. Zeng, et al., What makes a good retriever for open-domain question answering?, Transactions of the Association for Computational Linguistics 12 (2024) 1–19.

[21] Goover, Advancements in retrieval-augmented generation, https://seo.goover.ai/report/202502/go-public-report-en-d8f4e00f-9e56-4651-bc66-274977ebd6d2-0-0.html, 2025.

[22] CelerData, Latest developments in retrieval-augmented generation, https://celerdata.com/glossary/latest-developments-in-retrieval-augmented-generation, 2025.

[23] Chitika, Rag for code generation: Automate coding with ai & llms, https://www.chitika.com/rag-for-code-generation/, 2025.

[24] S. Solutions, Trends in active retrieval augmented generation: 2025 and beyond, https://www.signitysolutions.com/blog/trends-in-active-retrieval-augmented-generation, 2025.

[25] Y. Liu, Z. Wang, Y. Zhou, W. Zhang, Y. Y. Wang, Leveraging llm and user feedback to improve retrieval-augmented generation when question and answer domains shift, in: Proceedings of the 2024 Annual Meeting of the Association for Computational Linguistics (ACL), 2024. URL: https://openreview.net/forum?id=S034bjikkf.

[26] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://doi.org/10.1007/s10579-020-09502-8. doi:10.1007/s10579-020-09502-8.

[27] B. R. Chakravarthi, Detection of homophobia and transphobia in YouTube comments, International Journal of Data Science and Analytics 18 (2024) 49–68. URL: https://doi.org/10.1007/s41060-023-00400-0. doi:10.1007/s41060-023-00400-0.

[28] M. G. Yigezu, O. Kolesnikova, G. Sidorov, A. F. Gelbukh, Transformer-based hate speech detection for multi-class and multi-label classification., in: IberLEF@ SEPLN, 2023.

[29] J. J. Andrew, Judithjeyafreeda@ lt-edi-2023: Using gpt model for recognition of homophobia/-transphobia detection from social media, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, 2023, pp. 78–82.

[30] B. R. Chakravarthi, A. Hande, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, How can we detect Homophobia and Transphobia? Experiments in a multilingual code-mixed setting for social media governance, International Journal of Information Management Data Insights 2 (2022) 100119. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000623. doi:https://doi.org/10.1016/j.jjimei.2022.100119.

[31] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785/.

[32] M. Farouk, Measuring sentences similarity: A survey, Indian Journal of Science and Technology (2019). doi:10.17485/ijst/2019/v12i25/143977.

[33] UKPLab, Sentencetransformers documentation, 2024. Available at https://www.sbert.net.

[34] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260. URL: https://aclanthology.org/2022.acl-demo.25. doi:10.18653/v1/2022.acl-demo.25.