Cost-Effective Hate Speech Detection in Portuguese Using Lightweight LLMs

Mauro Cardoso
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
INESC-ID Lisboa, Portugal

Eugénio Ribeiro
INESC-ID Lisboa, Portugal
ISTAR, Instituto Universitário de Lisboa (ISCTE-IUL), Portugal

Fernando Batista
Instituto Universitário de Lisboa (ISCTE-IUL), Portugal
INESC-ID Lisboa, Portugal

Abstract

Addressing Online Hate Speech (OHS) in Portuguese faces significant challenges, including a lack of comprehensive, annotated data and classification frameworks, effective detection tools, and the high computational cost of Large Language Models (LLMs). This work evaluates the efficiency and competitiveness of seven smaller, more accessible LLMs (ranging from 3B to 27B parameters) through zero-shot classification with structured instructions, leveraging the expert-annotated test dataset and annotation scheme of the *kNOwHATE* project. Findings indicate that models such as *Mistral Small 3.2 24B*, *Gemma 3 27B*, and *Phi-4* achieve competitive performance, balancing precision and recall, whilst remaining computationally affordable. While excelling in identifying direct and indirect hate speech, out-group derogation, and specific target groups, the models faced challenges in detecting subtle rhetorical devices and emotions. Analysis of the *Cohen's Kappa* and F1-scores revealed moderate agreement with human annotators, highlighting the potential of optimised models to democratise OHS detection in resource-constrained settings, alongside the need for further refinement to bridge the gap in human-level nuance and improve generalisation. This study aims to broaden knowledge regarding the detection of OHS in Portuguese, by demonstrating a viable, competitive, and low-cost approach that does not rely on large-scale infrastructure or expensive proprietary models.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Human-centered computing → Social network analysis; Computing methodologies → Language resources; Applied computing → Sociology

Keywords and phrases Hate Speech Detection, Zero-shot Classification, Lightweight Language Models

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.9

Funding This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: `https://doi.org/10.54499/UID/50021/2025`), UID/PRR/50021/2025 (DOI: `https://doi.org/10.54499/UID/PRR/50021/2025`), and UID/04466/2025 (DOI: `https://doi.org/10.54499/UID/04466/2025`).

1 Introduction

The automatic detection of Online Hate Speech (OHS) has become a critical challenge in the age of social media, particularly in multicultural contexts such as Portuguese-speaking communities ([9]). Recent studies highlight the complexity of this multifaceted phenomenon, both in terms of its definition and its manifestation on social media platforms, given that this type of content is predominantly expressed through indirect and subtle language rather

than overt aggression, and it is often masked by discursive strategies and negative emotions. This type of discourse is primarily aimed at marginalised or minority communities, and has been systematically exploited by political actors as a strategy for polarisation and electoral mobilisation, intensifying social tensions and undermining democratic values ([18, 20, 6, 4, 21]). While Large Language Models (LLMs) have revolutionised the automatic detection of hateful content, they often struggle with slang, vulgar language, or context that may not be inherently toxic, and performance is highly influenced by their capability, purpose, techniques used and how prompts are designed.

Research on hate speech detection in English is well-established [27, 13], however, there's a shortage of publicly available labelled and annotated datasets in Portuguese, which poses a significant challenge to the development, refining, evaluation, and deployment of robust detection systems by general practitioners, limiting the applicability of state-of-the-art models ([26, 5]). Annotated data, being the cornerstone of training and evaluating supervised machine learning algorithms, allows models to capture the semantic nuances and context of human language. In this case, quality takes precedence over quantity: even small datasets, if balanced and carefully labelled by experts, can significantly improve model performance ([31, 7, 26]). Though essential, data annotation can be subjective, as annotators' perceptions vary depending on their identity and experiences ([10]), and success depends not only on the use of robust metrics to evaluate model performance, but also on the influence of annotators in the annotation process ([31, 9, 19]). Despite these challenges, the adoption of LLMs is further constrained by computational and financial costs due to the large number of parameters, and the ongoing need for specialised fine-tuning to handle tasks more efficiently, making their practical application unsustainable, particularly for non-profit organisations, small-scale platforms, or researchers in resource-constrained environments ([5, 2, 9, 1]). Moreover, the scope of analysis in common computational approaches remains limited when compared to the in-depth, systematic examination of OHS provided by specialised studies in Portuguese ([18, 29, 30, 28, 14]). Unlike the prevailing binary approaches, this study adopts the multidimensional labels defined in these works to capture the nuances of hate speech that are frequently overlooked, encompassing discourse types (direct vs. indirect), target communities (racialised groups, migrants, LGBTI+), discrimination processes, rhetorical devices, discursive strategies, fallacies, and emotions. In this context, modern classification methodologies, such as zero-shot classification using structured prompts, are emerging as promising alternatives, enabling the potential and generalisation capabilities of smaller or optimised language models to be exploited, reducing costs and providing a viable alternative with competitive performance ([19, 15, 9, 26, 5, 3]).

This work aims to fill these gaps by extending the analysis and evaluation of OHS components in Portuguese, testing and comparing low-cost language models in a predefined parameters range, offering an accessible and replicable methodology with competitive results. The following research questions are addressed:

- How do small-scale language models, chosen for their low computational resource requirements, perform in detecting OHS components in Portuguese using zero-shot classification with prompts?
- What is the effectiveness and trade-offs of these small-scale language models compared to human annotators?

This document is organised as follows: Section 1 introduces the study, and specifies the scope of the study. Section 2 is a comprehensive review of relevant prior work. Section 3 describes the research methodology, including the data and language models selection,

prompt design, introduces the multi-dimensional classification framework and the evaluation methodologies. Section 4 discusses the results and provides a detailed analysis of the findings. Finally, Section 5 summarises the main findings of the study, discusses its limitations and potential implications, and provides directions for future research.

2 Related Work

This section introduces relevant literature for the study, particularly in the Portuguese context, providing an initial understanding of the context of OHS, and then addressing relevant methodological approaches.

2.1 Understanding Online Hate Speech Content in Portuguese

Based on the *kNOwHATE: kNOwing online HATE speech*[1] project and associated research ([18, 29, 30, 28, 14]), OHS is seen as biased, harmful language that spreads or supports hate and violence against individuals or groups based on their social identity. From a social psychology perspective, OHS is an intergroup issue that reflects bias, often through negative, hostile comments aimed at out-groups. It can also involve seeing the out-group as a threat to the in-group's interests. OHS manifests in two primary forms: direct and indirect expressions, where direct expressions are characterised by explicitly biased and depreciative language, often involving slurs and provocative insults, and in the other hand, indirect hate speech is more common and complex, avoiding explicit depreciatory terms in favour of implicit meanings that must be deduced pragmatically by the audience. OHS often conveys harmful messages through subtle yet effective means, resorting to discursive strategies such as stereotyping, dehumanisation, and the articulation of threats, each reinforcing prejudiced assumptions or instilling fear. Rhetorical devices like irony, sarcasm, metaphors, hyperboles, and rhetorical questions further amplify these messages, often masking their harmful intent while undermining automated detection. Additionally, logical fallacies, such as calls to action or appeals to fear are used to justify hostility or incite harmful measures against targeted groups. Together, these strategies and devices normalise and propagate harmful ideologies within online spaces.

In Portugal, OHS is shaped by social, historical, and cultural factors, where research identifies four main groups targeted: racialised communities (including Afro-descendants and Roma), migrants, and LGBTI+ individuals. The way hate speech manifests against these groups is influenced by Portugal's distinct national narratives: Lusotropicalism, the idea that Portuguese colonisers were uniquely skilled at peaceful intercultural relations, downplaying systemic racism; Anti-Roma Sentiment, centring on stereotypes of criminality, laziness, and supposed resistance to society integration; Sexual Prejudice, targeting LGBTI+ communities through sexual stigma and the perception that they violate traditional in-group values, often triggering expressions of moral disgust and anger. When analysed on Portuguese social media platforms, using expert-annotated datasets (*YouTube*: 24,739 entries; *Twitter/X*: 29,758 entries) and a rigorous annotation process, including inter-annotator agreement measures, results reveal that discriminatory discourse is predominantly indirect, and the primary mechanism of discrimination involves the derogation of out-groups, manifested through stereotyping, threats, and dehumanisation, often amplified by emotions such as hate and anger. Discursive strategies vary according to the targeted community, with irony and

[1] `https://knowhate.eu/`

metaphors emerging as prevalent tactics, LGBTI+ communities are the primary targets on *Twitter/X*, while afrodescendant and roma communities face the most discrimination on *YouTube*. These findings highlight the urgent need to assess the scope and effectiveness of automated detection of OHS in the above contexts.

2.2 Overview on Automated Detection of Online Hate Speech

The automated detection of OHS and toxic content on social media sits at the crossroads of Natural Language Processing (NLP) and computational sociolinguistics. This field has evolved from simple dictionary-based approaches to advanced transformer models and LLMs. This progress has produced a shift from static, lexicon-based methods toward dynamic, context-aware architectures capable of more sophisticated semantic reasoning ([5, 19, 3]).

In this context, while fine-tuned Bidirectional Encoder Representations from Transformers (BERT) models perform well on data similar to their training, LLMs are more resilient to changes in data distribution, where in tests across different datasets, they tend to maintain higher F1-scores than BERT models, which can struggle when faced with unfamiliar data [8, 5]. These models have introduced novel classification capabilities that were largely unattainable with earlier techniques, most notably through zero-shot and few-shot detection of harmful content ([8, 9, 19, 3, 5]). Unlike traditional classifiers that require extensive task-specific training data, like annotated datasets, which its scarce in many languages ([8, 19, 31]), LLMs can identify toxic content by leveraging their broad pre-existing knowledge through natural language instructions, or prompts, opening the possibility of in-context learning, where models adapt to specific linguistic nuances simply by being provided with step-by-step thinking and a few illustrative examples in the input prompt ([2, 8, 5, 9, 19]). Furthermore, LLMs have fundamentally altered the role of the model from a simple classifier to an annotator or expert rationale generator, where research demonstrates that LLMs like *GPT-4* can outperform human crowd-workers in content moderation tasks, such as labelling toxic content, providing a scalable solution to the data scarcity problem in low-resource languages ([19, 3, 4]). Portuguese language relates to this scarcity, despite significant current progress. Even with advanced models, high-quality annotated data remains crucial, as it directly shapes model performance during both training and evaluation phases, and success depends on the precision of these annotations and the use of reliable evaluation metrics ([5, 10, 31]). However, this process depends on trained experts to define hate speech standards, though their judgments can still be influenced by sociocultural and personal biases. For example, an annotator's background can shape perceptions: someone of African descent might interpret aggressive language differently, and female annotators may more often flag gender-based insensitivity than male annotators ([10]).

Despite their cutting-edge status, LLMs are not a universal solution for OHS, requiring far more computational and financial resources than encoder-only models, which can sometimes be more effective for specific tasks, especially on social media ([8, 5]). Performance also depends heavily on prompt design and parametrisation, where harmless text can still be misclassified, particularly when it includes group identifiers or swear words ([8, 19]). This is where LLMs with smaller parameters come into play, typically ranging from 3B parameters upward, and within the zero-shot learning paradigm, they present a complex trade-off between computational efficiency and classification robustness. Though, the field has favoured large proprietary decoder models, recent studies show that compact, often open-source models can match their performance in toxic content detection, especially when fine-tuned or guided with optimised prompts for domain-specific tasks ([9, 19, 3]).

2.3 Zero-Shot Classification for Hate Speech Detection in Portuguese

Zero-shot classification looks promising for detecting hate speech in Portuguese, especially when interpretability and adaptability matter. Still, its practical use must account for computational demands, pointing toward smaller models as a way to balance effectiveness, cost, and efficiency.

Large scale decoder models as well as optimised models such as *GPT-4o* mini, demonstrate high resilience to data distribution shifts, often outperforming fine-tuned encoder models like *BERTimbau* in cross-dataset scenarios, offering sophisticated understanding to offensive content, providing natural language rationales for their decisions ([8, 9, 19]). Other smaller or optimised models, such as *LLaMA 3.1 8B*, *Phi-3-Mini (3.8B)*, and *Falcon (7B)*, offer a more accessible alternative for local deployment. While *LLaMA 3.1* 8B in its original zero-shot state may show lower F1-scores compared to larger models, it achieves high recall, although at the cost of precision ([8, 9]). When optimised through iterative prompt refinement, these 8B models can match or surpass the precision of encoder models like *BERTimbau*, and the performance of *GPT-4o* mini.

Studies show that the effectiveness of zero-shot classification depends on two key layers of prompts: the instruction and the context. A well-structured prompt typically integrates three core elements – system instruction, which defines the model's role and task boundaries; Contextual enhancement, which provides relevant keywords to guide multilingual models; Task specificity, where a balance must be struck between narrowly focused prompts (which yield high recall but lower precision) and broader prompts (which better balance precision and recall by including terms like aggression or offence) ([8, 9, 5, 2]). Template design often involves simple formats like "This text is" or "Classify this text as [option 1], [option 2], or [option 3].", a binary response "yes" or "no.", or advanced setups combining system messages (e.g., "You are performing text analysis") with user messages that include the specific question and the target text ([26, 8, 5]). For instructional styles, one approach is the *Vanilla* method, which directly asks whether content belongs to a class (e.g., "Is the following comment hateful?"). Another method is Chain-of-Thought (CoT), requiring the model to reason step-by-step before providing a label. A further technique is Role-Play, where the model is assigned a specific identity, such as a community moderator. Additionally, Definition-based prompts, which provide formal definitions of target classes to reduce ambiguity [15, 19, 2]. To accurately evaluate OHS detection in Portuguese, a robust approach is essential, one that accounts for the language's subjectivity and the scarcity of large annotated datasets. Performance is primarily measured using macro-averaged F1-scores to ensure fair comparison across imbalanced classes, while precision and recall help reveal biases ([8, 15]). Standardised frameworks, datasets and metrics such as *Cohen's Kappa*, *Fleiss' Kappa*, and *Krippendorff's* alpha, provide consistent benchmarks and inter-rater reliability, testing robustness through controlled cases involving implicit hate, negation, and non-hateful slurs. ([15, 19, 8, 5, 31, 18]).

3 Research Methodology

Having established the theoretical and practical challenges of OHS detection in Portuguese, this section outlines the data, models selection and prompt design rationale, the multi-dimensional framework, and lastly, the evaluation methodology.

3.1 Data Selection

As introduced in Section 2, this study uses the *Twitter/X* and *Youtube* test dataset made available by the project *kNOwHATE: kNOwing online HATE speech* ([17, 18, 29, 30, 28, 14]). The full dataset, collected between 2021 and 2022, comprises approximately 37,000 geolocated tweets originating from Portugal. It was gathered using a lexical-based approach, guided by a predefined list of terms tied to target communities, and to refine the dataset, it was applied a filtering strategy for ambiguous terms; for example, the word "preto" (which can mean "black" or serve as a slur) was only retained when it appeared alongside offensive or derogatory language. The dataset was meticulously annotated by a trained interdisciplinary team of social psychologists and linguists, who assessed each comment across 47 variables. These included the type of speech (offensive, direct or indirect hate speech,), targets (specific marginalized communities), discrimination processes (motivations such as out-group derogation or zero-sum perceptions), discursive and rhetorical features (stereotypes, dehumanization, threats, irony, metaphors, and logical fallacies like call to action or appeal to fear), and emotions (hate, anger, disgust, or fear). All variables were binary and non-mutually exclusive. Annotation reliability, measured using *Krippendorff's alpha*, revealed stronger agreement for objective categories like target identification than for subjective categories like subtle irony or specific emotions.

Due to the terms of the grant agreement with the European Union, it has been classified as sensitive and is therefore not publicly accessible. However, researchers can access a subset of the this golden set used for experimental test, via the Open Science Framework (OSF) ([17]), resulting in a total of 1630 posts with 62 columns and 47 labels. Figure 1 presents the distribution of positive (1) and negative (0) values of 29 labels in the context of OHS, filtered for classification and analysis ([18]), where the definitions of these labels will be explained in Section 3.3. The visualisation highlights a marked under-representation of positive label classes, initially suggesting significant challenges for classification and evaluation tasks.

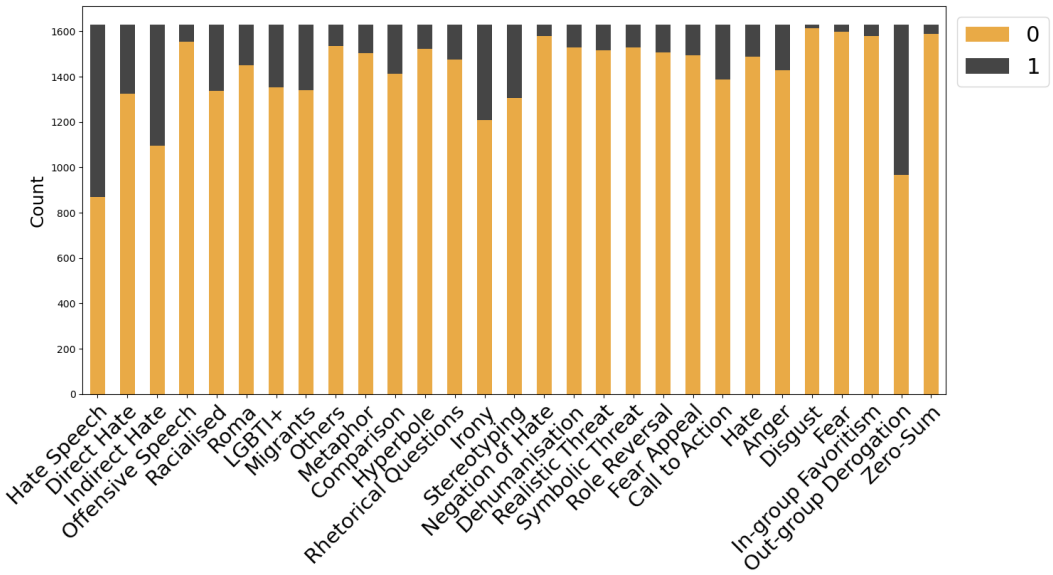

Figure 1 Selected Labels Distribution of the Twitter(X) and YouTube test data.

3.2 Language Model Selection and Activity Overview

A wide range of language models is now available, creating a dynamic and diverse ecosystem with varying strategies, efficiency levels, and specialisations. From this ecosystem, seven models were chosen based on their accessibility on OpenRouter, technical characteristics, provider uptime performance, reputation of the model's owner, and most importantly, the low cost. Table 1 presents some metrics related to the models selected and an overview of model usage within this study, in which we highlight the diverse parameter scale coverage among the selected language models. This distribution highlights the availability of both compact and mid-sized models, which are particularly relevant for resource-constrained environments, where the proprietary models (*Gemini 2.0 Flash Lite*) stands out from the open source remaining models, due to its undisclosed parameter count, suggesting a potential trade-off between performance and transparency. While input/output pricing varies (as is common in the industry, where outputs are often more expensive than inputs), all selected models allowed for cost optimisation depending on the provider, with highly competitive rates, at no more than $0.08 per million input tokens and no more than $0.30 per million output tokens. Parameter sizes range from 3B to 27B, influencing both capability and cost. Request volumes are relatively consistent across models, ranging from 47k to 60k per 1M tokens. *Llama 3.2 3B Instruct* emerges as the most economical, and *Gemma 3 27B* the most expensive. Models like *Gemini 2.0 Flash Lite* and *Mistral Small 3.2* offer a balance between cost and advanced features, while *Phi-4* is constrained by a shorter context window.

Table 1 Language models selected and activity overview for online hate speech detection based on cost efficiency, context length, and model capacity.

Model	Parameters	Context Len.	I/O Price*	Cost, Tokens, Requests
Google: Gemini 2.0 Flash Lite ([16])	NA	1.05M	$0.075/$0.30	$0.410, 2.54M, 48k
Google: Gemma 3 27B ([11])	27B	131k	$0.08/$0.16	$1.126, 6.18M, 50k
Meta: Llama 3.1 8B Instruct ([22])	8B	131k	$0.016/$0.021	$0.161, 6.69M, 47k
Meta: Llama 3.2 3B Instruct ([23])	3B	80k	$0.003/$0.006	$0.348, 6.74M, 48k
Mistral: Mistral Small 3.2 24B ([25])	24B	131k	$0.075/$0.20	$0.466, 5.82M, 48k
Mistral: Mistral Nemo ([24]	12B	128k	$0.02/$0.04	$0.1268, 5.79M, 47k
Microsoft: Phi-4 ([12])	14B	16.4	$0.065/$0.14	$0.5276, 8.14M, 60k

*Price per 1M tokens for input/output, depending on provider (OpenRouter, April 2026).

3.3 Multi-Dimensional Classification Framework

In order to expand and evaluate the ability of models to detect and analyse OHS components in Portuguese, it was necessary to establish a robust framework. This framework is based on the *kNOwHATE: kNOwing online HATE speech* project ([18, 29, 30, 28, 14, 17]), which provides both an annotation scheme and comprehensive guidelines that are central to this study. These elements form the theoretical and methodological foundation for the classification system employed in the selected models, where Tables 2 and 3 show this framework, which is based on a curated selection of 29 labels, aiming to simplify the scope of the classification approach.

These labels were derived from multiple analytical dimensions and definitions, as specified in the annotation scheme and refined through subsequent research [18, 17]. Each label is accompanied by its guiding questions and operational definitions, illustrating how the classification framework systematically integrates target groups, rhetorical devices, hate speech types, emotional markers, and discursive strategies into a cohesive analytical structure.

Table 2 Multi-Dimensional Framework for Zero-Shot Hate Speech Classification: Hate Speech Types, Target Groups, Rhetorical Devices, and Fallacies. (Part I).

Label	Question	Definition
	Hate Speech Types	
Hate Speech	Does this text contain hate speech?	Speech that contains incitement, insults or discrimination based on social identity, whether direct or indirect.
Direct Hate	Does this text express Direct Hate Speech?	Explicit and aggressive expression of hate against a group based on its social identity.
Indirect Hate	Does this text express Indirect Hate Speech?	Implicit or covert expression of hate, using subtleties or irony to avoid direct accusation.
Offensive Speech	Does this text express Offensive Speech (not necessarily hate)?	Aggressive or insulting language with an individual focus, not necessarily based on social group.
	Target Groups	
Racialised	Does this text mention or focus on racialised communities?	Hate speech targeting groups in processes of racialisation, attributing a constructed racial identity that often leads to social exclusion.
Roma	Does this text mention or focus on the Roma (gypsy) community?	Hate speech targeting racialised groups identified through the stigmatising term "gypsy" or similar.
Migrants	Does this text mention or focus on migrants?	Hate speech targeting individuals due to their nationality or migratory origin. Often related to xenophobia.
LGBTI+	Does this text mention or focus on LGBTI+ people?	Hate speech targeting sexual orientation, gender identity/expression, or sex characteristics. Encompasses homophobia, biphobia, transphobia, and intersexphobia.
Others	Does this text mention or focus on other minority groups outside the main categories (Racialised, Roma, Migrants, LGBTI+)?	Hate speech directed at groups or communities not explicitly included in the other categories (Racialised, Roma, Migrants, LGBTI+).
	Rhetorical Devices	
Metaphor	Does this text use the rhetorical device of Metaphor?	Transfer of attributes between distinct concepts with the aim of devaluing or dehumanising the target group.
Comparison	Does this text use the rhetorical device of Comparison?	Association of a group with another through analogies or comparisons, aiming to depreciate or devalue.
Hyperbole	Does this text use the rhetorical device of Hyperbole?	Use of intentional exaggeration to provoke emotional reactions or reinforce negative perceptions of the target group.
Rhetorical Questions	Does this text use Rhetorical Questions?	Questions that do not require an answer, but aim to reinforce a negative opinion or stereotype.
Irony	Does this text use the rhetorical device of Irony?	Indirect and ambiguous expression used to mock or humiliate, often in the form of sarcasm or humour.
	Fallacies	
Fear Appeal	Does this text use the rhetorical device of Fear Appeal?	Strategy that uses fear to persuade, suggesting negative outcomes if no action is taken.
Call to Action	Does this text use the rhetorical device Call to Action?	Explicit or implicit summons to act against the target group or a situation portrayed as negative.

Table 3 Multi-Dimensional Framework for Zero-Shot Hate Speech Classification: Discursive Strategies, Discrimination, and Emotions (Part II).

Label	Question	Definition
	Discursive Strategies	
Stereotyping	Does this text use Stereotyping?	Attribution of fixed and negative characteristics to a group based on incorrect generalisations.
Negation of Hate	Does this text use the strategy of Negation of Hate?	Strategy of presenting one's own discourse as legitimate and non-hateful, even when transgressing social norms.
Dehumanisation	Does this text use Dehumanisation?	Representation of the target group as less human, facilitating hostility or violence against it.
Realistic Threat	Does this text use the strategy of Realistic Threat?	Perception of concrete threat to the resources, security, or well-being of the dominant group.
Symbolic Threat	Does this text use the strategy of Symbolic Threat?	Perception of threat to the values, beliefs, or identity of the dominant group.
Role Reversal	Does this text use Role Reversal?	Presenting the dominant group as the victim and the marginalised group as the aggressor.
	Discrimination	
In-group favouritism	Does this text describe the form of discrimination known as in-group favouritism (an explicit preference for members of one's own group)?	An explicit preference for members of one's own group, even at the expense of impartiality.
Out-group Derogation	Does this text describe the form of discrimination known as 'out-group derogation' (the devaluation or hostility towards members of other groups)?	Belittling or hostility directed at external groups.
Zero-Sum	Does this text reflect a form of zero-sum discrimination (favouring one's own group and disparaging the out-group)?	An explicit preference for members of their own group and hostility towards external groups.
	Emotions	
Hate	Does this text express the emotion of Hate?	Intense aversion and desire to eliminate or harm a group perceived as evil or immutable.
Anger	Does this text express Anger?	Negative emotion motivated by perception of injustice, often with a desire for confrontation or change.
Disgust	Does this text express Disgust?	Moral or physical rejection of a group, often associated with ideas of contamination or devaluation.
Fear	Does this text express Fear?	Emotion associated with the perception of threat or danger (physical or symbolic) posed by the target group.
Guilt	Does this text express Guilt?	Recognition of moral responsibility for transgressions committed by the in-group.

This framework is especially valuable in zero or few shot classification tasks, where models must generalise from definitions and examples to identify OHS across diverse contexts, without prior exposure to specific instances. Such an approach directly supports the goals of initiatives like the kNOwHATE project, which seeks to develop culturally sensitive and effective methods for OHS detection and prevention, particularly within the Portuguese linguistic and cultural context.

3.4 Classification and Evaluation Approach

For this classification approach, prompts were carefully designed to determine whether a given comment contained a single label related to OHS, grounded in prior research, which shows that model performance tends to decline as the number of classes increases ([3]). The prompts followed a question–statement pair logic, where each text was submitted to a model with responses restricted to binary options – *"Yes"* or *"No"* to minimize ambiguity.

Each prompt consisted of three key components: a direct question to guide the model's analysis, a clear definition outlining the criteria, and the text for analysis. An example of the structure is given as follows:

Prompt Design Components

Question: "Does this text mention or focus on ..?"
Definition: "Hate speech targeting groups in processes of .. "
Text for analysis: "I hate this people... ."

Algorithm 1 shows the classification process applied, where it involves iterating over each dataset and text instance, while querying each language model with the prompt components to generate the predictions along with the relevant evaluation metrics. The metrics are F1-score, which harmonises precision and recall, false positive (FP) and false negative (FN) rates, and *Cohen's Kappa* which measures agreement beyond chance (Models Prediction vs Human Annotation).

Algorithm 1 Classification Approach for Hate Speech Detection.

Require:
Datasets $D = D_1, D_2, \ldots, D_n$ with annotations for hate speech.
Models $L = L_1, L_2, \ldots, L_m$ to evaluate.
Ensure:
Performance scores for each model on each dataset.
for each dataset $D_i \in D$ **do**
  for each text instance $t \in D_i$ **do**
    for each model $L_j \in L$ **do**
      Preprocess text if necessary.
      Create a prompt for model L_j.
      Send prompt to model L_j and obtain response r_{ij}.
      Parse and validate r_{ij} to get predicted label p_{ij}.
      Compare true labels with predictions.
      Compute metrics:
        Accuracy, Precision, Recall, F1-score, Cohen's Kappa, TP, FP, TN, FN.
    end for
  end for
end for
Join results.

4 Result Analysis

The results presented in this section are based on the methodology and multidimensional classification framework established in the previous sections, providing a comprehensive evaluation and analysis of the lightweight language models selected for the detection of OHS in Portuguese.

4.1 Models Performance Analysis

The interpretation of the results demands a cautious approach, given the unbalanced nature of the test data and the diversity of models/labels evaluated. The radar charts in Figure 2 provides the starting point for the analysis, comparing the performance of the models selected across five metrics: accuracy, precision, recall, *Cohen's Kappa*, and F1-score. The display is divided into three groups, each with its own legend showing the respective models, to make the analysis clearer.

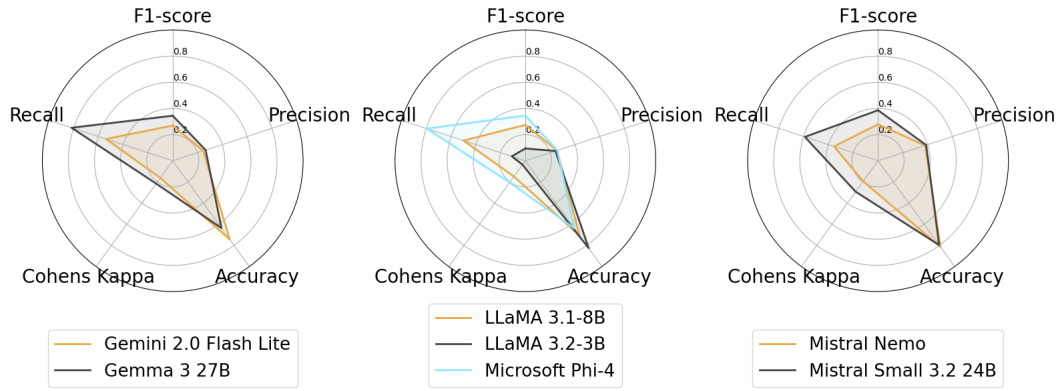

Figure 2 Overview of the models' average performance.

Overall, the models exhibit similar behaviour: although they are able to identify the majority of cases and maintain a good overall accuracy rate and recall, they suffer from a lack of precision. *LLaMA 3.2-3B*, for example, has the highest overall accuracy rate, but is the least balanced and consistent. *Mistral Small 3.2 24B*, on the other hand, stands out as the most robust, delivering a much more balanced performance than other equally robust competitors such as *Gemma 3 27B*, *Mistral Nemo* and *Phi-4*.

The heat maps in Figure 3 present a deeper comparative analysis of F1-score and *Cohen's Kappa* metrics, revealing distinct patterns in the models' ability to detect OHS. The LLaMA-3.2-3B model consistently shows lower F1-scores across most labels, particularly for positive instances, despite moderate results in "Stereotyping" (0.37) and "Irony" (0.42). As the most affordable but smallest model, its limited performance suggests it cannot fully capture complex OHS nuances without further refinement. Consequently, applying a strict exclusion criterion, where models failing to achieve an F1-score above 0.15 across more than 70% of labels are discarded to avoid sub-baseline noise. This model is excluded from further analysis in favour of better-performing alternatives.

The *Mistral Small 3.2 24B*, *Gemma 3 27B* and *Phi-4* models stand out for their superior performance in both metrics. Specifically, they achieve high F1-scores for labels such as "LGBTI+", "Out-group Derogation", and "Direct Hate Speech", indicating a robust global capacity for detecting hateful content. Although these models imply higher costs, their

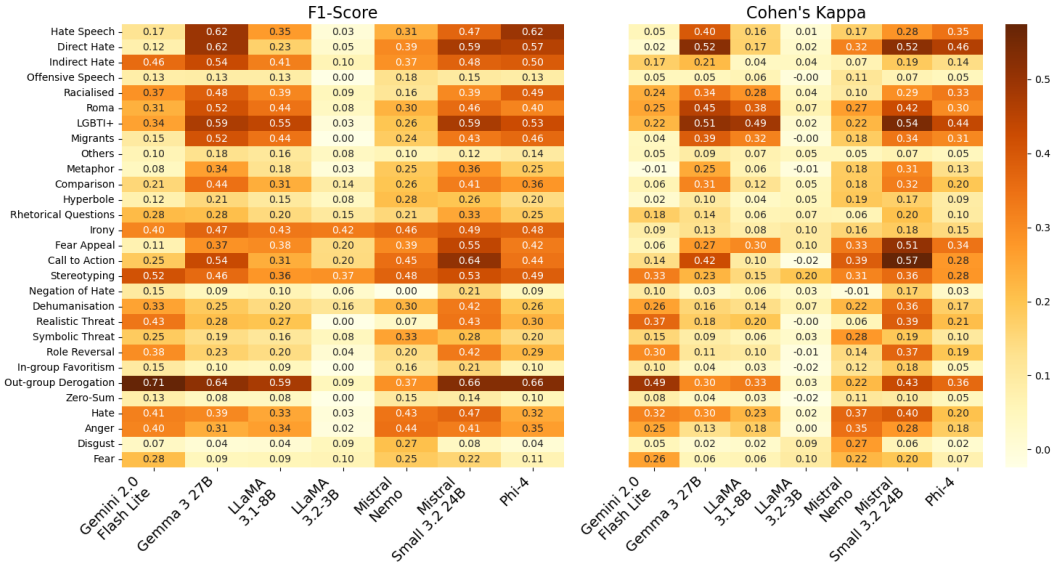

Figure 3 F1-score and *Cohen's Kappa* heatmap for each model by OHS labels.

performance is considered justified and proportional to the complexity and scale of the information processed. The *Gemini 2.0 Flash Lite*, despite its extensive context window, shows only moderate performance relative to the other models. This observation suggests that an excessive context window may not be critical for this specific task, as the model does not significantly outperform others. A notable pattern emerges in the models' ability to generalise across various labels, though difficulties persist in handling less frequent nuances or subclasses, such as "Denial of Hate", "Disgust", "Zero-Sum", and "In-group Favouritism". Additionally, the models exhibit challenges in interpreting labels like "Fear", "Metaphor", "Hyperbole", and "Offensive Speech", which are often camouflaged forms of OHS.

Despite these limitations, the models show a robust capacity for identifying direct and indirect hate speech, aligning with findings from previous research in the field. The F1-scores across models range from moderate to high, while *Cohen's Kappa* values tend to be lower. This discrepancy suggest that, although the models can effectively detect OHS categories, inter-annotator agreement or internal consistency may be constrained. Such a trend underscores the complexity of achieving both high accuracy and reliable human-model agreement in this domain. For applications prioritising cost-effectiveness, the *Mistral Nemo*, *Gemini 2.0 Flash Lite*, and *Llama 3.1 8B* models emerge as the most balanced options, although they would likely require some fine-tuning. These models are particularly suitable for scaling to large datasets without substantially compromising performance, making them ideal for projects with more budget constraints. For scenarios demanding higher performance – within a controlled budget – the *Mistral Small 3.2 24B*, *Gemma 3 27 B* and *Phi-4* models stand out as a valid choice. Although they require a higher budget for scaling, the results demonstrate significant value, justifying the investment for tasks where precision and reliability are paramount. Given that, in this case, fine-tuning would likely yield high-quality results. Considering the confusion matrix shown in Figure 4, which presents even more detailed analysis of the top-performing models, *Gemma 3 27B* and *Phi-4* show high sensitivity, detecting a large number of positive cases (true positives).

Label	Mistral Small 3.2 24B				Mistral Nemo				Google Gemma 3 27B				Microsoft Phi-4			
	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN	TP	FP	TN	FN
Hate Speech	248	45	823	514	144	22	846	618	388	104	764	374	422	183	685	340
Direct Hate	155	62	1262	151	84	40	1284	222	199	141	1183	107	193	177	1147	113
Indirect Hate	285	367	727	251	201	338	756	335	388	516	578	148	360	555	539	176
Offensive Speech	70	804	749	7	37	292	1261	40	75	972	581	2	71	911	642	6
Racialised	92	88	1250	200	30	43	1295	262	173	251	1087	119	202	336	1002	90
Roma	54	4	1447	125	34	13	1438	145	105	120	1331	74	109	254	1197	70
Migrants	99	74	1265	192	45	34	1305	246	181	229	1110	110	181	310	1029	110
LGBTI+	122	13	1341	154	43	7	1347	233	155	96	1258	121	153	143	1211	123
Others	11	75	1459	85	9	76	1458	87	47	375	1159	49	50	545	989	46
Metaphor	43	69	1434	84	35	118	1385	92	109	405	1098	18	104	607	896	23
Comparison	90	127	1286	127	45	81	1332	172	168	372	1041	49	174	572	841	43
Hyperbole	94	514	1009	13	84	415	1108	23	103	788	735	4	99	790	733	8
Rhetorical Questions	134	534	942	20	101	708	768	53	148	770	706	6	146	889	587	8
Irony	370	722	485	53	317	639	568	106	404	893	314	19	401	851	356	22
Fear Appeal	89	95	1398	48	53	83	1410	84	119	386	1107	18	98	230	1263	39
Call to Action	151	81	1306	92	81	38	1349	162	206	313	1074	37	206	495	892	37
Stereotyping	246	363	941	80	192	287	1017	134	297	673	631	29	283	553	751	43
Negation of Hate	20	118	1461	31	0	6	1573	51	51	1034	545	0	51	1030	549	0
Dehumanisation	85	216	1313	16	70	297	1232	31	98	592	937	3	89	495	1034	12
Realistic Threat	55	83	1432	60	4	3	1512	111	99	503	1012	16	89	387	1128	26
Symbolic Threat	89	450	1077	14	49	141	1386	54	97	802	725	6	98	756	771	5
Role Reversal	56	86	1421	67	20	61	1446	103	101	670	837	22	86	382	1125	37
In-group Favouritism	13	60	1520	37	16	130	1450	34	37	691	889	13	39	669	911	11
Out-group Derogation	423	203	763	241	165	52	914	499	521	446	520	143	507	368	598	157
Zero-Sum	36	441	1147	6	27	296	1292	15	40	879	709	2	41	764	824	1
Hate	111	219	1270	30	78	141	1348	63	130	392	1097	11	130	549	940	11
Anger	185	506	922	17	108	177	1251	94	200	876	552	2	195	731	697	7
Disgust	16	381	1233	0	3	3	1611	13	16	847	767	0	16	834	780	0
Fear	14	82	1516	18	14	68	1530	18	27	535	1063	5	28	465	1133	4

Figure 4 Top performant models: Confusion matrix metrics by label.

However, this broad coverage results in a clear tendency towards over-classification, as evidenced by the high rates of false positives in almost all categories. In contrast, the Mistral family is showing a more cautious approach, *Mistral Small 3.2 24B* and the *Mistral Nemo* offer high accuracy, with a very low rate of false positive errors compared to other models. The cost of this selectivity is an increase in missed cases (false negatives), particularly in complex categories such as "Out-group Derogation" and "Hate Speech". Finally, categories that require a nuanced reading between the lines, such as "Irony" and "Indirect Hate Speech", prove to be a common challenge. In these cases, all models struggle to distinguish the underlying intent, resulting in a delicate balance between correct and incorrect classifications.

5 Conclusions and Future Work

This study suggests that lightweight language models ranging from 12B to 27B parameters, such as *Mistral Small 3.2 24B*, *Gemma 3 27B* and *Phi-4*, serve as computationally efficient and cost-effective alternatives for detecting OHS in Portuguese. By using a zero-shot classification approach within a multi-dimensional framework adapted from the *kNOwHATE* project, these models achieve competitive performance, finding a necessary balance between classification robustness and operational affordability.

Regarding the first research question (RQ1), the findings shows that smaller-scale models can achieve competitive performance for OHS detection in low resource environments ([19, 9]), particularly when combined with techniques like *zero-shot* classification optimised through prompt engineering and structured definitions. While these models excel at identifying direct and indirect hate speech and specific target groups, such as *LGBTI+*, *Racialised communities*, and *Roma*, the detection of nuanced rhetorical devices – including irony, hyperbole, and metaphor – remains a persistent challenge. This difficulty often leads to over-classification and a higher rate of false positives, which highlights the ongoing need for more in-depth semantic reasoning to accurately analyse the underlying intent. In relation to the second research question (RQ2), the alignment between model outputs and human annotations, as measured by *Cohen's Kappa*, reveals a nuanced performance landscape. Across most labels, the models achieved moderate agreement with the expert-annotated set, though consistently below the inter-annotator agreement observed among human experts. Specifically, for example models in general demonstrated relatively high recall and accuracy, aligning well with the majority of expert annotations for direct and indirect hate speech, out-group derogation, and target group identification. However, their precision lagged behind human annotators, particularly for ambiguous or culturally specific expressions.

Several limitations must be acknowledged, notably the reliance on specific sample datasets and the presence of class imbalance, which may have skewed performance metrics by inflating recall while masking precision deficiencies. Furthermore, the zero-shot paradigm lacks the domain-specific fine-tuning necessary to adapt to evolving slang and the dynamic nature of hateful discourse. Not including different and adaptive suggestion techniques, and the constraints of shorter context windows may also have limited the models' ability to capture long-range dependencies in more complex textual structures.

Future research should prioritise enhancing dataset representativeness and exploring hybrid strategies that integrate zero/few-shot and fine-tuning approaches. To achieve this, it is essential to address model bias and deploy these frameworks in diverse environments, including political contexts and community-led monitoring, thereby ensuring more ethical, transparent, and culturally sensitive OHS detection solutions for the Portuguese language.

References

1 Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, Fatemeh Jamshidi, Mostafa Rahgouy, Nilanjana Raychawdhary, Daniela Marghitu, and Cheryl Seals. Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access*, 13:20871–20892, 2025. doi:10.1109/ACCESS.2025.3532397.

2 Gabriel Assis, Annie Amorim, Jonnathan Carvalho, Daniel de Oliveira, Daniela Vianna, and Aline Paes. Exploring Portuguese Hate Speech Detection in Low-Resource Settings: Lightly Tuning Encoder Models or In-Context Learning of Large Models? In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 301–311. Association for Computational Linguistics, 2024. URL: https://aclanthology.org/2024.propor-1.31/.

3 Basel Barakat and Sardar Jaf. Beyond Traditional Classifiers: Evaluating Large Language Models for Robust Hate Speech Detection. *Computation*, 13(8), 2025. doi:10.3390/computation13080196.

4 Mauro Cardoso, Eugénio Ribeiro, and Fernando Batista. Portuguese Far-Right Discourse on Social Media: Insights from Topic Modeling. In *14th Symposium on Languages, Applications and Technologies (SLATE 2025)*, volume 135 of *Open Access Series in Informatics (OASIcs)*, pages 12:1–12:16. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/OASIcs.SLATE.2025.12.

5 Amanda et al. da Silva Oliveira. Toxic Speech Detection in Portuguese: A Comparative Study of Large Language Models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 108–116. Association for Computational Linguistics, 2024. URL: https://aclanthology.org/2024.propor-1.11/.

6 Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515, 2017. doi:10.1609/icwsm.v11i1.14955.

7 J. Angel Diaz-Garcia and Joao Paulo Carvalho. A Literature Review of Textual Cyber Abuse Detection Using Cutting-Edge Natural Language Processing Techniques: Language Models and Large Language Models. *WIREs Data Mining and Knowledge Discovery*, 15(3):e70029, 2025. doi:10.1002/widm.70029.

8 Amanda Oliveira et al. How Good Is ChatGPT For Detecting Hate Speech In Portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103, Porto Alegre, RS, Brasil, 2023. SBC. doi:10.5753/stil.2023.233943.

9 Amanda Oliveira et al. Toxic Text Classification in Portuguese: Is LLaMA 3.1 8B All You Need? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 57–66. SBC, 2024. doi:10.5753/stil.2024.245416.

10 Amit Das et al. Investigating Annotator Bias in Large Language Models for Hate Speech Detection. *ArXiv*, 2024. doi:10.48550/arXiv.2406.11109.

11 Gemma Team et al. Gemma 3 technical report, 2025. doi:10.48550/arXiv.2503.19786.

12 Microsoft et. al. Phi-4 technical report. *arXiv*, 2024. arXiv:2412.08905v1.

13 Zahra Safdari Fesaghandis and Suman Kalyan Maity. Multilingual hate speech detection and counterspeech generation: A comprehensive survey and practical guide. *arXiv preprint arXiv:2603.19279*, 2026. doi:10.48550/arXiv.2603.19279.

14 Pedro Fialho, Ricardo Ribeiro, Fernando Batista, Gil Ramos, António Fonseca, Sérgio Moro, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. *Counter Hate Speech Detection in Youtube Conversations*, pages 94–105. Springer Nature Switzerland, 2025. doi:10.1007/978-3-031-73997-2_9.

15 Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. Can Prompting LLMs Unlock Hate Speech Detection across Languages? A Zero-shot and Few-shot Study. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 413–425, Vienna, Austria, August 2025. Association for Computational Linguistics. URL: https://aclanthology.org/2025.woah-1.39/.

16 Google DeepMind. Gemini 2.0 flash lite: Documentation and technical overview. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash-lite, 2024. Access: April 2026.

17 Rita Guerra, Fernando Batista, Paula Carvalho, and Ricardo Ribeiro. Data (Golden set/Test files), November 2025. doi:10.17605/OSF.IO/MPEJ3.

18 Rita Guerra, Paula Carvalho, Catarina Marques, Margarida Carmona, Rodrigo Sarroeira, Fernando Batista, Ricardo Ribeiro, António Fonseca, Sérgio Moro, and Cláudia Silva. Unpacking online hate speech in Portuguese social media: a social-psychological and linguistic-discursive approach. *Humanities and Social Sciences Communications*, 12:1709, 2025. doi:10.1057/s41599-025-05392-9.

19 Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. Harnessing Artificial Intelligence to Combat Online Hate: Exploring the Challenges and Opportunities of Large Language Models in Hate Speech Detection. *arXiv*, 2024. doi:10.48550/arXiv.2403.08035.

20 Tiago Lapa and Branco di Fátima. Hate Speech Among Security Forces in Portugal. *Communication Library*, 7:277–293, 2023. doi:10.25768/654-916-9.

21 Luca Manucci. Portuguese Populism: People, Parties, and Politics. *Análise Social*, 59(2 (251)):2–11, 2024. doi:10.31447/202200.

22 Meta. Llama 3.1: Model cards & prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/, 2024. Access: April 2026.

23 Meta. Llama 3.2: Model cards & prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/, 2024. Access: April 2026.

24 Mistral AI. Mistral-nemo: Architecture and performance report. https://mistral.ai/news/mistral-nemo, 2024. Access: April 2026.

25 Mistral AI. Mistral 3: Release notes and benchmarks. https://mistral.ai/news/mistral-small-3, 2025. Access: April 2026.

26 Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.woah-1.6.

27 Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, June 2021. doi:10.1007/s10579-020-09502-8.

28 Catarina Pontes, António Fonseca, Sérgio Moro, Fernando Batista, Ricardo Ribeiro, Catarina Marques, Paula Carvalho, Cláudia Silva, and Rita Guerra. *Unveiling Patterns of Hate Speech in the Portuguese Sphere: A Social Network Analysis Approach*, pages 70–81. Springer Nature Switzerland, 2025. doi:10.1007/978-3-031-73997-2_7.

29 Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. A comprehensive review on automatic hate speech detection in the age of the transformer. *Social Network Analysis and Mining*, 14:204, 2024. doi:10.1007/s13278-024-01361-3.

30 Gil Ramos, Fernando Batista, Ricardo Ribeiro, Pedro Fialho, Sérgio Moro, António Fonseca, Rita Guerra, Paula Carvalho, Catarina Marques, and Cláudia Silva. Leveraging Transfer Learning for Hate Speech Detection in Portuguese Social Media Posts. *IEEE Access*, 12:101374–101389, 2024. doi:10.1109/ACCESS.2024.3430848.

31 Natalia Umansky, Maël Kubli, Ana Kotarcic, Laura Bronner, Selina Kurer, Philip Grech, Dominik Hangartner, Fabrizio Gilardi, and Karsten Donnay. Improving hate speech detection with large language models. *European Journal of Political Research*, pages 1–12, January 2026. doi:10.1017/S1475676525100546.