

SarcasTraps: Effect of Sarcasm on Turkish Hate Speech Detection

Niyazi Ahmet Metin Dilara Keküllüoğlu

Sabancı University

Istanbul, Türkiye

{ahmet.metin,dilara.kekulluoglu@sabanciuniv.edu}

Abstract

Standard hate speech detection datasets often emphasize explicit harmful language, but directed harm can also be expressed indirectly, without clearly hateful wording. Sarcasm is one such form of indirect expression, as it can convey harmful intent through mock praise, exaggeration, humor, or apparent politeness. In this paper, we introduce SarcasTraps, a targeted external test set for analyzing the effect of sarcasm on Turkish hate speech detection. SarcasTraps contains 479 entries annotated along two dimensions: hate speech and sarcasm. We train classifiers using BERTurk and XLM-RoBERTa on two existing Turkish hate speech datasets and evaluate them on SarcasTraps, using it strictly as an external test set. We also evaluate GPT-4.1 and Llama-3.3-70B in a zero-shot setting on SarcasTraps. Results show that sarcasm substantially affects hate speech detection. Fine-tuned encoder models lose recall on sarcastic hate, while zero-shot LLMs maintain higher recall but suffer substantial precision drops on sarcastic examples. These findings show that both fine-tuned hate speech detectors and zero-shot LLMs struggle when sarcasm is involved, and SarcasTraps provides quantitative evidence of this challenge through a targeted Turkish test set. Due to the presence of sensitive and hateful language, SarcasTraps will be shared upon request for academic research purposes. **Warning: This paper has examples from the dataset which contains offensive and hateful content towards groups of people.**

1 Introduction

Understanding sarcasm is difficult, as even its definition is contested. Work in pragmatics repeatedly notes that the boundary between sarcasm and irony is elusive, and broader theories of irony have not settled on a single stable definition. (Dyner, 2017; Kapogianni, 2021). Some sources take sarcasm as a subclass of irony (Leggitt and Gibbs, 2000), and some count them as separate notions (Ling and

Klinger, 2016). In general, sarcasm is described in terms of its overlap with irony rather than in isolation, in both, meanings inferred beyond the literal utterance, and there is a need to recover what a speaker intends rather than what the sentence literally says. Prior work on Turkish also suggests that the term “sarcasm” has no exact equivalent in Turkish; instead, several near-equivalents are used in everyday language. In Turkish literary tradition, sarcasm is most closely associated with “hiciv” (Banasik-Jemielniak et al., 2022). Within pragmatics, sarcasm could be treated not as a simple gap between literal and intended meaning. Sarcasm is also a communicative resource since figurative language is used to accomplish discourse goals, and sarcastic irony is often used to complain or criticize implicitly while making the speaker appear less rude or unfair (Roberts and Kreuz, 1994; Jorgensen, 1996).

In Natural Language Processing, sarcasm has most commonly been studied as a text classification problem, where the goal is to determine whether a given text is sarcastic. Early work focused largely on English social media and microblog texts, and later extended to other genres such as sarcastic headlines and crowdsourced corpora (Barbieri et al., 2014; Ptáček et al., 2014; Shrikhande et al., 2020; Oraby et al., 2016). More recent work has also considered intended sarcasm, where labels are provided by the authors of the sarcastic texts themselves (Oprea and Magdy, 2020). Beyond English, sarcasm and irony detection have been explored in languages such as Chinese and Arabic, as well as in multilingual settings including Arabic–English and Czech–English (Gong et al., 2020; Abu Farha and Magdy, 2020; Abu Farha et al., 2022; Ptáček et al., 2014). In Turkish, existing resources have appeared under partially overlapping labels such as irony, satire, and sarcasm: earlier datasets focused on Twitter or microblog-style texts (Taslioglu and Karagoz, 2017; Dülger, 2018; Karabaş and Dırı,

2020), while later work introduced *IronyTR*, Turkish satire datasets, and news-column datasets annotated for irony and sarcasm (Ozturk et al., 2021; Onan and Toçoğlu, 2020; Eser and Bilgin, 2025). Closest to our setting, Metin et al. (2026) introduced a context-aware Turkish sarcasm detection dataset from Ekşi Sözlük, using title-level contextual summaries to evaluate sarcasm detection in informal online discourse.

These studies provide the foundation for modeling sarcasm directly, but they also motivate asking how sarcasm interacts with other NLP tasks that depend on speaker intent. Thus, this study shifts the focus from sarcasm classification to investigating its broader effects; specifically, we examine whether sarcasm creates a hidden failure mode in the safety-critical domain of hate speech detection.

Hate speech detection is central to online safety, content moderation, and harmful language understanding (Schmidt and Wiegand, 2017), yet many evaluation settings primarily evaluate cases where harmful intent is expressed relatively directly, through explicit insults, slurs, or negative statements about a target group. (Waseem and Hovy, 2016; Davidson et al., 2017; Zampieri et al., 2019; ElSherief et al., 2021). Sarcasm challenges this assumption because it can separate the literal wording of a text from its intended social meaning. A sarcastic utterance may appear humorous, polite, exaggerated, or indirectly critical on the surface while still conveying contempt, exclusion, or hostility toward a protected group. This motivates our central question: Do hate speech detectors trained on standard Turkish datasets remain reliable when hateful intent is expressed sarcastically? This problem is especially relevant in Turkish online discourse because informal platforms frequently contain indirect, mocking, and culturally loaded expressions.

To address this question, we introduce SarcasTraps where we analyze the effect of sarcasm on Turkish hate speech detection. Rather than treating sarcasm as the target label, SarcasTraps annotates each instance along two dimensions: whether the text contains hate speech and whether it is sarcastic. Combining these annotations yields four conditions: non-hateful non-sarcastic, non-hateful sarcastic, hateful non-sarcastic, and hateful sarcastic. In total, SarcasTraps contains 479 samples. We use SarcasTraps as an external test set by training hate speech classifiers on existing Turkish hate speech datasets and evaluating how their predictions trans-

fer to these four classes.

We build SarcasTraps from Ekşi Sözlük (literally, “Sour Dictionary”), one of the largest user-generated discussion platforms in Türkiye (Ekşi Sözlük). The platform is organized around *titles*, where each title names a concept, event, person, phrase, or social issue, and users contribute short *entries* under that title. This structure produces a distinctive form of informal online discourse: entries are often subjective, culturally loaded, politically charged, humorous, or mocking, and their interpretation may depend on the shared topic established by the title. These properties make Ekşi Sözlük a suitable source for SarcasTraps, since it naturally contains cases where harmful group-level meaning can be expressed indirectly through sarcasm, irony, exaggeration, or mock praise.

2 Related Work

Hate speech detection has been widely studied as a supervised classification problem, especially over social media texts. Early work introduced Twitter-based datasets and feature-based models for identifying hateful or offensive content, using lexical and surface-level signals such as character n-grams, indicative words, and hate speech keywords (Waseem and Hovy, 2016; Davidson et al., 2017). Subsequent work expanded the task beyond binary hate detection by introducing large-scale abusive language datasets and more fine-grained annotation schemes for offensive content, including labels for offense type and target (Founta et al., 2018; Zampieri et al., 2019). Recent work has also explored domain-adapted transformer encoders such as HateBERT for abusive and hate speech detection (Caselli et al., 2021). However, recent work has noted that much hate speech detection research focuses on explicit or overt hate, leaving coded, indirect, or implicit forms less directly addressed (ElSherief et al., 2021). This concern has motivated recent evaluation settings for implicit and adversarial hate speech, where harmful meaning can be expressed without overt hateful wording (ElSherief et al., 2021; Hartvigsen et al., 2022). This line of work motivates our setting: if hate speech can be conveyed indirectly, then models trained on standard hate speech datasets may fail when group-directed hostility is expressed through non-literal forms such as sarcasm.

For Turkish, hate speech detection has been studied mostly through social media datasets and eval-

uation settings. Early Turkish resources include tweet-based corpora for offensive or abusive language detection, establishing initial baselines for harmful language identification in Turkish (Çöltekin, 2020). Mayda et al. (2021) introduce a publicly available Turkish hate speech dataset of tweets annotated as hate, offensive, or none, with additional subclass labels for hateful examples such as ethnic, religious, sexist, and political hate. More recent work scales Turkish hate speech detection to larger and more domain-oriented settings: Toraman et al. (2022) construct large-scale Turkish and English hate speech datasets and study cross-domain transfer, while Dehghan and Yanikoglu (2024) provides Turkish hate speech benchmarks across target domains such as refugees, Greeks, and Israel–Palestine-related discourse. Beyond post-level classification, Şeker et al. (2025) introduce HATECAT-TR, a span-annotated Turkish hate speech dataset that identifies hateful spans and categorizes them into fine-grained discourse types. These resources have substantially advanced Turkish hate speech detection, but they primarily model hate speech as post-level or span-level harmful language classification. In contrast, our work uses Turkish hate speech supervision to test whether such models remain reliable when hateful intent is expressed through sarcasm.

Recent work has increasingly examined how harmful language detection is affected when abuse or hate is expressed indirectly. Frenda et al. (2023b) investigates irony-aware models for abusive language detection, showing that incorporating irony and sarcasm information can improve the detection of abusive content, especially when harmful meaning is expressed indirectly. Frenda et al. (2023a) shows that harmful content may appear with a weakly offensive surface form. Their analysis supports the idea that abusive intent can rely on indirect cues like sarcasm or irony. Some corpus construction efforts have also moved beyond single-label hate speech annotation by including additional pragmatic and discourse-level dimensions, including sarcasm. For example, Friðriksdóttir et al. (2024) introduce Ice and Fire, an Icelandic blog-comment dataset annotated for multiple dimensions. Also, recent multi-dimensional social media datasets similarly annotate hate speech or offensive content together with related dimensions such as sarcasm, humor, stance, emotion, factuality, or spam (Gautam et al., 2020; Zaghouni et al., 2024; Pramanik et al., 2026).

3 SarcasTraps

To analyze the effect of sarcasm on Turkish hate detection, we curated the SarcasTraps dataset. All examples in SarcasTraps were collected from publicly accessible Ekşi Sözlük pages between April and June 2026.

Candidate titles were identified using Turkish search terms related to Greek, Jewish/Israeli, and immigrant/refugee discourse. We purposively inspected titles whose framing made group-directed hostility reasonably likely, rather than inspecting all keyword matches. For example, a topic-neutral title such as Syrian cuisine was not selected, whereas a title framed around an influx of Syrians could be inspected.

The target domains were selected as they frequently appear in Turkish online discussions and have been used in Turkish hate speech studies and benchmarks (Dehghan and Yanikoglu, 2024; Şeker et al., 2025; Mayda et al., 2021). The scope of SarcasTraps is limited to these domains. Entries whose hateful content targets only specific individuals are not included in the set. To protect privacy, individual person names occurring in the entries are masked using <PERSON_n>. In some discussions, users expressed hate and/or sarcasm in a specific sentence or phrase within a longer entry. In these cases, we extracted only the relevant portion containing the expression, rather than collecting the entire entry.

Three native Turkish annotators independently labeled each entry for sarcasm and hate speech. Two of the annotators are authors of this paper, while the third is an external annotator. The Krippendorff’s α over the independently assigned labels was 0.755 for hate speech and 0.659 for sarcasm. After this independent annotation stage, the annotators met to discuss disagreements. Consensus was reached for all but five entries, whose final labels were determined through majority voting. During adjudication, four entries were removed because they did not satisfy the scope of the test set, including entries that expressed hostility toward a specific individual rather than a group-level identity. The final test set contains 479 entries. The final distribution consists of 124 non-hateful non-sarcastic, 107 non-hateful sarcastic, 154 hateful non-sarcastic, and 94 hateful sarcastic entries.

Annotators were instructed to label an entry as sarcastic when its intended meaning differs from, reverses, or distorts its literal surface meaning. The

sarcasm label includes mock politeness, mock approval, deliberate distortion, rhetorical questions, irony, parody, feigned ignorance, or absurd exaggeration. Prior work, (Oraby et al., 2016), has shown that these forms frequently participate in sarcastic or ironic utterances, while also occurring in non-sarcastic contexts. These categories are not treated as separate annotation categories. The presence of any of these forms was treated as practical cues only when they supported a non-literal interpretation of the intended meaning.

Annotators were instructed to label an entry as hate speech if its intended meaning expresses a negative generalization, demeaning characterization, exclusionary stance, or hostile attitude toward a target group. The hate label was assigned only when the entry targeted a group-level identity or social category, rather than merely criticizing a specific individual. For sarcastic examples, annotators were instructed to reconstruct the intended proposition behind the sarcastic surface form before assigning the hate label. For example, if an entry appears to praise a target group in a sarcastic way and the intended interpretation is the opposite of that praise (e.g., a demeaning or hostile claim about the group), the example is labeled as hate. In contrast, sarcastic examples are labeled as non-hate only when the reconstructed intended meaning does not contain group-directed hostility, degradation, exclusion, or harmful generalization.

Examples from all four combinations of the hate speech and sarcasm labels are presented in Table 1. We report only English translations and partially mask profane expressions in the illustrative examples to reduce direct traceability and unnecessary exposure to harmful language. The underlying test set preserves the original wording, except that names of identifiable individuals are replaced with placeholders of the form <PERSON_n>.

The non-sarcastic examples provide a reference condition for separating general domain-transfer effects from the additional effect of sarcasm.

4 Experiments

The purpose of our experiments is to test whether hate speech detectors trained on Turkish hate speech datasets remain reliable when hateful intent is expressed through sarcasm. To this end, we trained encoder-only classifiers using two Turkish hate speech datasets to which we obtained access. We evaluated the trained models both on the corre-

sponding in-domain test sets and on the sarcastic and non-sarcastic subsets of SarcasTraps. We additionally evaluated two zero-shot large language models on SarcasTraps. The purpose of including multiple model configurations is not to establish a ranking across architectures, model scales, or training regimes. Instead, the models serve as complementary probes for examining whether changes across evaluation conditions are consistent across different system types. Accordingly, our analysis focuses on within-model performance differences rather than direct comparisons between fine-tuned encoders and zero-shot LLMs.

In the experiments, the models are evaluated as binary hate speech classifiers. That is, the target label is whether the input contains hate speech, while the sarcasm annotation is used only for analysis. SarcasTraps is used only as an external evaluation set, not for training or fine-tuning.

To train hate speech detectors for SarcasTraps test evaluation, we used two existing Turkish hate speech datasets as training resources: the Mayda et al. dataset (Mayda et al., 2021) and SU-HS (Dehghan and Yanikoglu, 2024). The Mayda et al. dataset was obtained from the authors’ public GitHub repository, where no explicit license was identified. The SU-HS dataset was obtained from the authors under a data-use agreement. Both datasets were used for academic research experiments, we cite the original creators and do not redistribute the datasets. We selected these datasets because they both cover the target domains represented in SarcasTraps, including refugees/Syrians, Greeks, and Jewish/Israel–Palestine-related hate speech. This allows us to test whether models trained with standard hate speech supervision transfer to sarcastically expressed hate in the same broad target domains. For Mayda et al., the original dataset contains three labels: hate, offensive, and none, and covers 22 target groups, including Syrian refugees, Greeks, and Jews. Since our experiments focus on binary hate speech detection, we excluded the offensive class and retained only the hate and none examples, resulting in 2,336 hate and 7,722 non-hate instances. We used examples from all 22 target groups in our experiments. As the dataset does not provide an official train/test split, we constructed a split where for each target group, we randomly sampled 10% of the examples as the test set and used the remaining 90% for training. This resulted in 9,051 training instances and 1,007 test instances. For SU-HS, we used the dataset’s

Title	Entry	H	S
Turkish and Greek soldiers swearing at each other	Made me laugh. Trash talk like this is better than actually fighting.	0	0
Do the Freemasons or the Jews rule the world?	Minibus drivers and taxi drivers rule the world.	0	1
A message for Afghans and Pakistanis	First of all, you are not our brothers. We don't have predatory, dishonorable brothers like you. Leave. Get the f*** out of our country.	1	0
The invasion of Caddebostan Beach by Syrians	Oh, but they are Ansar; they are our brothers. What's the big deal if they went into the sea in a chador, or if their men disturbed the peace of society? There is a war! They got tired while fleeing the war; shouldn't they enjoy the beach a little bit?	1	1

Table 1: Examples from the four conditions in SarcasTraps. H indicates hate speech and S indicates sarcasm.

provided binary train/test split without additional label filtering. The dataset covers four domains, including refugee-related hate speech, the Israel–Palestine conflict, and Greek-related hate speech. We used all available examples in these splits, resulting in 10,932 training instances and 2,743 test instances.

We trained separate encoder-based classifiers for each training dataset using BERTurk¹ (Schweter, 2020) and XLM-RoBERTa² (Conneau et al., 2020). During evaluation on SarcasTraps, the model input is formed as title + “ : ” + entry. We did not perform hyperparameter search and fine-tuned all encoder models for five epochs with a batch size of 16, maximum sequence length of 384, weight decay of 0.01, warmup ratio of 0.06. We used a learning rate of 4×10^{-5} for BERTurk and 2×10^{-5} for XLM-RoBERTa. Each encoder configuration was trained with five random seeds: 42, 123, 456, 789, and 1024. Encoder fine-tuning was implemented with PyTorch (Paszke et al., 2019) and the Hugging Face Transformers Library (Wolf et al., 2020). Evaluation metrics were computed with scikit-learn (Pedregosa et al., 2011). Fine-tuning was performed in Google Colab sessions, each using a single NVIDIA A100 GPU. Each SU-HS runs each consisted of 3,420 optimization steps, while each Mayda et al. runs each consisted of 2,830 steps. Across the four encoder configurations and five random seeds, this resulted in 20 fine-tuning runs. The total training time was approximately 1.8

GPU hours. After training, each classifier was first evaluated on the corresponding in-domain test set and then directly evaluated on SarcasTraps without any additional adaptation.

For decoder-only large language models, we performed zero-shot hate speech classification without fine-tuning. GPT-4.1 (OpenAI, 2025) and Llama-3.3-70B-Versatile (Meta, 2024) were evaluated using the same SarcasTraps instances. GPT-4.1 was accessed through the OpenAI API, and Llama-3.3-70B-Versatile was accessed through the Groq API.³ The models are prompted in Turkish and instructed to decide whether the given text contains hate speech. Hate speech is defined in the prompt as content that targets, demeans, threatens, or encourages violence against a person or group on the basis of identity-related attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or similar characteristics. The prompt constrains the output to a JSON object with a binary hate label, together with a short justification. We use deterministic decoding with temperature 0. The full prompts are provided in Appendix A.

For all models, we report aggregate hate speech detection performance on the in domain test sets and hate-class precision, recall, and F1 on the non-sarcastic and sarcastic subsets of SarcasTraps. Since the purpose of SarcasTraps is to test whether hate speech detection systems behave differently when hate is expressed sarcastically, the main results are analyzed by comparing model behavior across non-sarcastic and sarcastic cases. Detailed performance patterns and error analyses are pre-

¹<https://huggingface.co/dbmdz/bert-base-turkish-128k-cased>

²<https://huggingface.co/xlm-roberta-base>

³<https://console.groq.com/docs/api-reference>

sented in Section 5.

5 Results & Analysis

In this section, we evaluate model behavior across two successive conditions. We first examine whether the fine-tuned encoder models transfer from their corresponding in-domain test sets to the non-sarcastic subset of SarcasTraps. We then use this non-sarcastic subset as a reference condition and compare it with the sarcastic subset to analyze the effect of sarcasm on hate speech detection.

Table 2 compares the encoder models’ performance on their corresponding in-domain test sets and on the non-sarcastic subset of SarcasTraps, denoted SarcasTraps (0). Zero-shot LLMs are not included in this comparison because they were not trained on either of the in-domain datasets and therefore do not have a corresponding in-domain evaluation condition.

The results from Table 2 indicate that the models do not undergo a general performance collapse when transferred to the non-sarcastic portion of the targeted external test set. We therefore use SarcasTraps (0) as the reference condition for examining the additional change associated with sarcasm. This comparison should not be interpreted as showing that the in-domain test sets and SarcasTraps (0) have identical difficulty or class distributions; its purpose is to establish that the models retain meaningful hate speech detection ability in the absence of sarcasm.

Table 3 compares performance on the non-sarcastic reference condition, SarcasTraps (0), and the sarcastic subset, SarcasTraps (1). Unlike the preceding analysis, this comparison includes the zero-shot LLMs because both conditions are available for all evaluated models.

The degradation associated with sarcasm is systematic across all four encoder configurations. On SarcasTraps (0), the encoder models obtain mean F1 scores between 0.740 and 0.828. On SarcasTraps (1), their mean F1 scores fall to between 0.390 and 0.483. The absolute F1 decrease is highly consistent across the configurations, ranging from 0.344 to 0.350.

The encoder models’ main limitation on sarcastic examples is low hate-class recall. Recall decreases from 0.655-0.818 on SarcasTraps (0) to 0.294-0.404 on SarcasTraps (1), an absolute reduction of 0.338-0.415. Moreover, recall remains below precision for every encoder configuration

on the sarcastic subset. This indicates that the fine-tuned encoders frequently fail to retrieve hateful examples when group-directed hostility is expressed through a sarcastic surface form. Because the pattern occurs under both encoder architectures and both training datasets, it is not specific to a single configuration.

The zero-shot LLMs exhibit a different error profile. GPT-4.1’s F1 decreases from 0.899 on SarcasTraps (0) to 0.670 on SarcasTraps (1), while Llama-3.3-70B’s F1 decreases from 0.846 to 0.620. Both models retain comparatively high hate-class recall on sarcastic examples: 0.819 for GPT-4.1 and 0.755 for Llama-3.3-70B. However, their precision decreases from 0.826 to 0.566 and from 0.744 to 0.526, respectively. Their comparatively high recall therefore comes with substantially more false-positive hate predictions on non-hateful sarcastic examples.

To better understand the two main error patterns observed in the quantitative results, we examined two groups of misclassified instances in detail. For the encoder models, we examined hateful-sarcastic instances misclassified as non-hateful. Encoder errors were determined from the majority prediction across the five random seeds. For the zero-shot LLMs, we examined non-hateful-sarcastic instances misclassified as hateful.

At least one encoder configuration missed 81 of the 94 hateful-sarcastic instances (86.2%), while all four configurations missed 41 instances (43.6%).

Qualitative inspection revealed recurring errors involving indirect hostility expressed through apparent approval or mock praise, as well as implied threats or references to violence. These errors were not limited to seed-sensitive boundary cases. Depending on the encoder configuration, 68.9-80.4% of the hateful-sarcastic false negatives were misclassified by all five seeds. Moreover, 23 instances were assigned the non-hate label by every seed of all four encoder configurations.

The LLMs showed the opposite tendency on non-hateful sarcasm. At least one LLM produced a false positive on 77 of the 107 non-hateful-sarcastic instances (72.0%), and both LLMs misclassified 46 of these instances. Qualitative inspection suggested that the LLMs often misclassified non-hateful sarcastic political comments and criticism as hate speech.

Model	In-domain Test				SarcasTraps (0), $n = 278$			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
BERTurk _{SU-HS}	0.831 $\pm$ 0.010	0.704 $\pm$ 0.017	0.684 $\pm$ 0.058	0.693 $\pm$ 0.030	0.817 $\pm$ 0.009	0.871 $\pm$ 0.016	0.787 $\pm$ 0.011	0.827 $\pm$ 0.007
BERTurk _{Mayda}	0.940 $\pm$ 0.004	0.886 $\pm$ 0.010	0.852 $\pm$ 0.017	0.869 $\pm$ 0.009	0.812 $\pm$ 0.020	0.838 $\pm$ 0.018	0.818 $\pm$ 0.042	0.828 $\pm$ 0.022
XLm-RoBERTa _{SU-HS}	0.823 $\pm$ 0.005	0.689 $\pm$ 0.016	0.673 $\pm$ 0.013	0.680 $\pm$ 0.007	0.778 $\pm$ 0.028	0.898 $\pm$ 0.020	0.675 $\pm$ 0.046	0.770 $\pm$ 0.034
XLm-RoBERTa _{Mayda}	0.916 $\pm$ 0.006	0.854 $\pm$ 0.009	0.771 $\pm$ 0.040	0.810 $\pm$ 0.019	0.746 $\pm$ 0.027	0.854 $\pm$ 0.034	0.655 $\pm$ 0.041	0.740 $\pm$ 0.031

Table 2: Encoder performance on the corresponding in-domain test sets and the non-sarcastic subset of SarcasTraps. Precision, recall, and F1 are computed for the hate speech class. Results are reported as mean $\pm$ standard deviation over five random seeds.

Model	SarcasTraps (0), $n = 278$				SarcasTraps (1), $n = 201$			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
BERTurk _{SU-HS}	0.817 $\pm$ 0.009	0.871 $\pm$ 0.016	0.787 $\pm$ 0.011	0.827 $\pm$ 0.007	0.630 $\pm$ 0.026	0.692 $\pm$ 0.034	0.372 $\pm$ 0.063	0.483 $\pm$ 0.060
BERTurk _{Mayda}	0.812 $\pm$ 0.020	0.838 $\pm$ 0.018	0.818 $\pm$ 0.042	0.828 $\pm$ 0.022	0.597 $\pm$ 0.007	0.604 $\pm$ 0.014	0.404 $\pm$ 0.045	0.483 $\pm$ 0.031
XLm-RoBERTa _{SU-HS}	0.778 $\pm$ 0.028	0.898 $\pm$ 0.020	0.675 $\pm$ 0.046	0.770 $\pm$ 0.034	0.623 $\pm$ 0.013	0.746 $\pm$ 0.024	0.294 $\pm$ 0.030	0.421 $\pm$ 0.032
XLm-RoBERTa _{Mayda}	0.746 $\pm$ 0.027	0.854 $\pm$ 0.034	0.655 $\pm$ 0.041	0.740 $\pm$ 0.031	0.536 $\pm$ 0.020	0.509 $\pm$ 0.035	0.317 $\pm$ 0.009	0.390 $\pm$ 0.014
GPT-4.1 Zero Shot	0.878	0.826	0.987	0.899	0.622	0.566	0.819	0.670
Llama-3.3-70B Zero Shot	0.802	0.744	0.981	0.846	0.567	0.526	0.755	0.620

Table 3: Hate speech detection performance on the non-sarcastic and sarcastic subsets of SarcasTraps. Precision, recall, and F1 are computed for the hate speech class. Encoder results are reported as mean $\pm$ standard deviation over five random seeds. Zero-shot LLM results are from deterministic decoding with temperature 0.

6 Conclusion

In this paper, we presented SarcasTraps: Effect of Sarcasm on Turkish Hate Speech Detection. In the set, each example is annotated for both hate speech and sarcasm, yielding four conditions that separate literal hate, non-hateful sarcasm, and sarcastically expressed hate. This design allows us to evaluate not only whether models detect hate speech in general, but also whether they remain reliable when harmful intent is expressed indirectly.

Experiments with BERTurk and XLm-RoBERTa classifiers trained on existing Turkish hate speech datasets, evaluated across five random seeds, as well as zero-shot LLMs, show that sarcasm exposes a systematic weakness in current hate speech detection systems. Fine-tuned encoder-based models tend to miss sarcastically expressed hate, while zero-shot LLMs retain higher recall on sarcastic hate but also produce more false positives on non-hateful sarcastic examples. These results indicate that sarcasm affects hate speech detection in different ways across model families and that standard evaluation settings can obscure these failures.

Overall, SarcasTraps demonstrates that sarcasm is not only a phenomenon to be detected on its own. It is also a factor that can affect model behavior in safety-critical tasks. It shows that robust hate speech detection requires models to distinguish

not only harmful from non-harmful language, but also literal harmfulness from indirectly expressed group-directed hostility.

7 Limitations

SarcasTraps is designed as a targeted test set for evaluating the effect of sarcasm on Turkish hate speech detection systems. It contains 479 examples, which limits the scale of the quantitative evaluation. This size is partly due to the difficulty of finding naturally occurring examples that simultaneously involve hate speech and sarcasm. While hateful language and sarcastic language are both common in informal online discourse, examples where the two interact clearly enough to support reliable annotation are much rarer. For this reason, the results should be interpreted as evidence from a targeted test set.

Another limitation is the class distribution of the set. Because Ekşi Sözlük is organized around titles, data collection required identifying relevant titles and then selecting entries under those titles. To obtain enough hateful examples, we intentionally inspected titles where hate speech was more likely to occur, such as discussions involving socially or politically salient target groups. As a result, SarcasTraps contains more hateful than non-hateful examples. This imbalance reflects the targeted purpose of the dataset, but it also means that the set does not represent the natural prevalence of hate

speech or sarcasm.

Future work can extend SarcasTraps by adding more examples and covering additional target domains beyond the ones considered in this study. A larger and more diverse benchmark would allow more fine-grained analyses. Another important direction is the development of sarcasm-aware hate speech detection systems. The results in this paper suggest that models trained on standard hate speech datasets are not sufficient for robustly handling sarcastically expressed hate, so future systems may benefit from sarcasm detection signals, or multi-task learning setups that jointly consider sarcasm and hate speech.

Finally, future work should examine whether context improves hate speech detection in sarcastic cases. Prior work showed that title-level context can help sarcasm detection in Ekşi Sözlük-style discourse. A natural next step is to investigate whether similar title-level or discussion-level context can also improve the detection of hateful intent when it is expressed indirectly through sarcasm. Such work could clarify whether failures on sarcastic hate are caused primarily by missing pragmatic cues in the entry itself, or by the absence of broader contextual information about the title, event, or target group being discussed.

Because SarcasTraps focuses on hate speech and informal online discourse, it contains sensitive and harmful language that is not censored in order to preserve the original content.

8 Ethical Considerations

This study aims to support research on sarcasm-aware hate speech detection in Turkish online discourse. Because SarcasTraps focuses on hate speech and informal online discourse, it contains sensitive and harmful language that is not censored in order to preserve the original content.

SarcasTraps is constructed from naturally occurring entries on Ekşi Sözlük and is intended solely as a research resource for analyzing how sarcasm affects hate speech detection systems. The dataset should not be used to amplify hateful content or to target individuals or communities as the set itself may cause harm if redistributed without restrictions, or used outside of its intended research purpose.

All examples were collected by the authors. The examples were independently annotated by three native Turkish speakers: the two authors and an

Turkish MSc student from the same laboratory who is not an author. The annotators are native Turkish speakers who are familiar with Ekşi Sözlük discourse and the goals of the project. The annotators were not recruited through a crowdsourcing platform and did not receive payment. We do not report protected demographic attributes of annotators, beyond task-relevant information such as native Turkish proficiency and familiarity with Ekşi Sözlük discourse. Before collection and annotation, they were informed that the data would contain hateful, offensive, and disturbing content. During annotation, they focused only on two specific dimensions: whether an entry contains hate speech and whether it is sarcastic. They did not evaluate the political correctness, factual accuracy, moral acceptability, or ideological position of the entries.

Because the set was constructed from a public source, and the collection followed the platform’s robots.txt restrictions, no formal ethics review was obtained for this study. We do not release the set as a public download; instead, access will be provided only upon request for academic-only purposes. The dataset contains only the title, entry content, hate label, and sarcasm label. We do not collect or release usernames, timestamps, profile information, or other metadata that could be used to identify Ekşi Sözlük authors to reduce privacy risks. We did not obtain individual consent from Ekşi Sözlük users, as the entries were collected from publicly accessible pages. We acknowledge that public data does not always translate to consent for research (Fiesler and Proferes, 2018). As suggested in (Fiesler and Proferes, 2018), we are not mass gathering entries from specific people. The titles and entries in the set do not include any identifying information. We masked individual person names appearing in the entries, but all the other content remains uncensored. Due to sensitive nature of the data, SarcasTraps will be distributed only upon academic request under research-only terms.

Large language models were used as zero-shot classifiers in the experiments; they were not used to generate SarcasTraps entries or assign the annotations. We used ChatGPT 5.6 and Grammarly to check spelling and text flow after drafting our own version, and incorporated suggestions for more appropriate wording.

References

- Ibrahim Abu Farha and Walid Magdy. 2020. [From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset](#). In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, Marseille, France. European Language Resource Association.
- Ibrahim Abu Farha, Steven Wilson, Silviu Oprea, and Walid Magdy. 2022. [Sarcasm detection is way too easy! an empirical comparison of human and machine sarcasm detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5284–5295, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Natalia Banasik-Jemielniak, Piotr Kałowski, Büşra Akkaya, Aleksandra Siemieniuk, Yasemin Abayhan, Duygu Kandemirci-Bayız, Ewa Dryll, Katarzyna Brannowska, Anna Olechowska, Melanie Glenwright, Maria Zajączkowska, Magdalena Rowicka, and Penny M. Pexman. 2022. [Sarcasm use in turkish: The roles of personality, age, gender, and self-esteem](#). *PLOS ONE*, 17(11):1–16.
- Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. [Modelling sarcasm in Twitter, a novel approach](#). In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, Baltimore, Maryland. Association for Computational Linguistics.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. [HateBERT: Retraining BERT for abusive language detection in English](#). In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.
- Çağrı Çöltekin. 2020. [A corpus of Turkish offensive language on social media](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. [Automated hate speech detection and the problem of offensive language](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.
- Somaiyeh Dehghan and Berrin Yanikoglu. 2024. [Multi-domain hate speech detection using dual contrastive learning and paralinguistic features](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11745–11755, Torino, Italy. ELRA and ICCL.
- Oğuzhan Dülger. 2018. [Türkçe metinlerde ironi tespiti \(irony classification in turkish text\)](#). In *Proceedings of the 12th Turkish National Software Engineering Symposium (UYMS 2018)*, volume 2201 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Marta Dynel. 2017. [Academics vs. american scriptwriters vs. academics: A battle over the etic and emic “sarcasm” and “irony” labels](#). *Language & Communication*, 55:69–87. Conversational humour: Spotlight on languages and cultures.
- Ekşi Sözlük. 1999. [Ekşi sözlük](#).
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Murat Eser and Metin Bilgin. 2025. [Irony and sarcasm detection in turkish texts: A comparative study of transformer-based models and ensemble learning](#). *Applied Sciences*, 15(23).
- Casey Fiesler and Nicholas Proferes. 2018. [“participant” perceptions of twitter research ethics](#). *Social Media+ Society*, 4(1):2056305118763366.
- Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. [Large scale crowdsourcing and characterization of twitter abusive behavior](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).
- Simona Frenda, Viviana Patti, and Paolo Rosso. 2023a. [Killing me softly: Creative and cognitive aspects of implicitness in abusive language online](#). *Natural Language Engineering*, 29(6):1516–1537.
- Simona Frenda, Viviana Patti, and Paolo Rosso. 2023b. [When sarcasm hurts: Irony-aware models for abusive language detection](#). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, page 34–47, Berlin, Heidelberg. Springer-Verlag.
- Steinunn Rut Friðriksdóttir, Annika Simonsen, Atli Snær Ásmundsson, Guðrún Lilja Friðjónsdóttir, Anton Karl Ingason, Vésteinn Snæbjarnarson, and Hafsteinn Einarsson. 2024. [Ice and fire: Dataset on sentiment, emotions, toxicity, sarcasm, hate speech, sympathy and more in Icelandic blog comments](#). In

- Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 73–84, Torino, Italia. ELRA and ICCL.
- Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. [#metoo: Multi-aspect annotations of tweets related to the metoo movement](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):209–216.
- Xiaochang Gong, Qin Zhao, Jun Zhang, Ruibin Mao, and Ruifeng Xu. 2020. [The design and construction of a Chinese sarcasm dataset](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5034–5039, Marseille, France. European Language Resources Association.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. [ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.
- Julia Jorgensen. 1996. [The functions of sarcastic irony in speech](#). *Journal of Pragmatics*, 26(5):613–634.
- Eleni Kapogianni. 2021. Types and definitions of irony. In Piotr Stalmaszczyk, editor, *The Cambridge Handbook of the Philosophy of Language*, Cambridge Handbooks in Language and Linguistics, pages 622–638. Cambridge University Press, Cambridge.
- Ahmet Karabaş and Banu Dırı. 2020. [Irony detection with deep learning in turkish microblogs](#). In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.
- John S Leggett and Raymond W Gibbs. 2000. Emotional reactions to verbal irony. *Discourse processes*, 29(1):1–24.
- Jennifer Ling and Roman Klinger. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *European semantic web conference*, pages 203–216. Springer.
- İslam Mayda, Yunus Emre Demir, Tuğba Dalyan, and Banu Dırı. 2021. [Hate speech dataset from turkish tweets](#). In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.
- Meta. 2024. [Llama 3.3: Model Cards and Prompt Formats](#). Meta Llama Documentation. Accessed: 2026-05-21.
- Niyazi Ahmet Metin, Sevde Yılmaz, Osman Enes Erdoğan, Elif Sude Meydan, Oğul Sümer, and Dilara Keküllüoğlu. 2026. [SarcasTürk: Turkish context-aware sarcasm detection dataset](#). In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 61–71, Rabat, Morocco. Association for Computational Linguistics.
- Aytuğ Onan and Mansur Alp Toçoğlu. 2020. [Satire identification in turkish news articles based on ensemble of classifiers](#). *Turkish Journal of Electrical Engineering and Computer Sciences*, 28(2):1086–1106.
- OpenAI. 2025. [GPT-4.1 Model](#). OpenAI API Documentation. Accessed: 2026-05-21.
- Silviu Oprea and Walid Magdy. 2020. [iSarcasm: A dataset of intended sarcasm](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.
- Shereen Oraby, Vrindavan Harrison, Lena Reed, Ernesto Hernandez, Ellen Riloff, and Marilyn Walker. 2016. [Creating and characterizing a diverse corpus of sarcasm in dialogue](#). In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–41, Los Angeles. Association for Computational Linguistics.
- Asli Umay Ozturk, Yesim Cemek, and Pinar Karagoz. 2021. [Ironytr: Irony detection in turkish informal texts](#). *International Journal of Intelligent Information Technologies*, 17(4):1–18.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.
- Souvik Pramanik, S. M. Riaz Rahman Antu, Shak Mohammad Abyad, Md. Ibrahim Khalil, and Md. Shahriar Hussain. 2026. [Multisoc-4d: A benchmark for diagnosing instruction-induced label collapse in closed-set llm annotation of bengali social media](#). *Preprint*, arXiv:2605.06940.
- Tomáš Ptáček, Ivan Habernal, and Jun Hong. 2014. [Sarcasm detection on Czech and English Twitter](#). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.
- Richard M. Roberts and Roger J. Kreuz. 1994. [Why do people use figurative language?](#) *Psychological Science*, 5(3):159–163.

Anna Schmidt and Michael Wiegand. 2017. [A survey on hate speech detection using natural language processing](#). In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Stefan Schweter. 2020. Berturk-bert models for turkish. *Zenodo*.

Hasan Kerem Şeker, Gökçe Uludoğan, Pelin Önal, and Arzucan Özgür. 2025. [HATECAT-TR: A hate speech span detection and categorization dataset for Turkish](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25568–25579, Suzhou, China. Association for Computational Linguistics.

Parnavi Shrikhande, Vikram Setty, and Dr. Ashish Sahani. 2020. [Sarcasm detection in newspaper headlines](#). In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 483–487.

Hande Taslioglu and Pinar Karagoz. 2017. [Irony detection on microposts with limited set of features](#). In *Proceedings of the Symposium on Applied Computing, SAC '17*, page 1076–1081, New York, NY, USA. Association for Computing Machinery.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. [Large-scale hate speech detection with cross-domain transfer](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Zeeraq Waseem and Dirk Hovy. 2016. [Hateful symbols or hateful people? predictive features for hate speech detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wajdi Zaghouni, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. [So hateful! building a multi-label hate speech annotated Arabic dataset](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055, Torino, Italia. ELRA and ICCL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. [Predicting the type and target of offensive](#)

[posts in social media](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

A Zero-shot LLM Prompt for Hate Speech Classification

We used the same Turkish prompt for zero-shot hate speech classification with GPT-4.1 and Llama-3.3-70B-Versatile. The models were instructed to return a binary JSON label and a short justification. The original Turkish system message is provided below.

A.1 Turkish System Message

Sen bir nefret söylemi (hate speech)
→ sınıflandırıcısın.
Kullanıcı sana Türkçe bir metin verecek.
Görevin: metnin nefret söylemi içerip
→ içermediğine karar vermek.

Nefret söylemi; bir kişiyi veya grubu ırk,
→ etnisite, din, cinsiyet,
cinsel yönelim, engellilik veya benzeri kimlik
→ özelliklerine dayanarak
hedef alan, aşağılayan, tehdit eden veya şiddete
→ teşvik eden içeriktir.

Yalnızca aşağıdaki JSON formatında yanıt ver, kod
→ bloğu veya ek metin OLMADAN:
{ "label": 1 | 0, "reason": "<kısa gerekçe>" }

1 = nefret söylemi var, 0 = nefret söylemi yok.
Emin değilsen en olası etiketi seç ve bunu
→ gerekçede belirt.

A.2 English Translation of System Message

You are a hate speech classifier.
The user will provide a Turkish text.
Your task is to decide whether the text contains
→ hate speech.

Hate speech is content that targets, demeans,
→ threatens, or encourages
violence against a person or group on the basis
→ of identity-related
attributes such as race, ethnicity, religion,
→ gender, sexual orientation,
disability, or similar characteristics.

Respond only in the following JSON format,
→ without code fences or extra text:
{ "label": 1 | 0, "reason": "<short reason>" }

1 = hate speech is present, 0 = hate speech is
→ not present.
If uncertain, choose the most likely label and
→ state this in the reason.

B User Prompt Template

For each SarcasTraps instance, we formed the input text by concatenating the title and entry content as

title + " : " + content, and inserted the resulting text into the following user prompt.

Aşağıdaki metni analiz et ve nefret söylemi
↔ içerip içermediğine karar ver.

Metin:
{{TEXT}}

Yalnızca şu JSON nesnesini döndür:
{ "label": 1 | 0, "reason": "<kısa gerekçe>" }

We provide the English translation below for clarity.

Analyze the following text and decide whether it
↔ contains hate speech.

Text:
{{TEXT}}

Return only the following JSON object:
{ "label": 1 | 0, "reason": "<short reason>" }

C LLM Output Parsing

For both models, we used JSON object response formatting and deterministic decoding with temperature 0. Outputs were parsed as JSON objects and the label was normalized to a binary value, where 1 indicates hate speech and 0 indicates no hate speech.