SarcasTürk: Turkish Context-Aware Sarcasm Detection Dataset

Niyazi Ahmet Metin **Sevde Yılmaz** **Osman Enes Erdoğdu**
Elif Sude Meydan **Oğul Sümer** **Dilara Keküllüoğlu**
Sabanci University
Istanbul, Türkiye
{ahmet.metin,sevdenaz.yilmaz,osmanerdogdu,elifsude,
osumer,dilara.kekulluoglu}@sabanciuniv.edu

Abstract

Sarcasm is a colloquial form of language that is used to convey messages in a non-literal way, which affects the performance of many NLP tasks. Sarcasm detection is not trivial and existing work mainly focus on only English. We present SarcasTürk, a context-aware Turkish sarcasm detection dataset built from Ekşi Sözlük entries, a large-scale Turkish online discussion platform where people frequently use sarcasm. SarcasTürk contains 1,515 entries from 98 titles with binary sarcasm labels and a title-level context field created to support comparisons between entry-only and context-aware models. We generate these contexts by selecting representative sentences from all entries under a title using summarization techniques. We report baseline results for a fine-tuned BERTurk classifier and zero-shot LLMs under both no-context and context-aware conditions. We find that BERTurk model with title-level context has the best performance with 0.76 accuracy and balanced class-wise F1 scores (0.77 for sarcasm, 0.75 for no sarcasm). SarcasTürk can be shared upon contacting the authors since the dataset contains potentially sensitive and offensive language.

1 Introduction

Sarcasm is a form of indirect expression in which the intended meaning diverges from the literal wording. It is commonly used to express criticism and to deliver praise in a playful or humorous way (Banasik-Jemielniak et al., 2022). Sarcasm can be hard to catch for humans (Pexman et al., 2019) and machines alike (Weitzel et al., 2016). However, it is an integral part of human communication (Gibbs, 2000), so automated sarcasm detection is vital for many NLP tasks such as sentiment analysis (Weitzel et al., 2016; Maynard and Greenwood, 2014). Currently, most sarcasm detection research is in English, with minimal work on languages such as Turkish.

In many online settings, sarcastic intent is shaped by shared topic knowledge and the ongoing theme of a discussion, so contextual understanding can be critical for interpretation. This makes sarcasm highly dependent on contextual cues and shared assumptions between the writer and the reader (Oprea and Magdy, 2020b). Prior work emphasizes that the same sentence can be intended ironically or non-ironically depending on contextual factors, and even human annotators request additional context to reliably infer ironic intent (Wallace et al., 2014). These properties make Turkish informal online discourse, such as Ekşi Sözlük, an especially relevant setting for studying sarcasm as a context-sensitive phenomenon.

Ekşi Sözlük ("Sour Dictionary") (Ekşi Sözlük) is one of the largest user-generated discussion platforms in Türkiye. It is an open-topic forum where users contribute messages (referred to as "entries"). Each entry belongs to a shared topic ("title"), and all entries under the same title form a discussion with the same shared topic. The language used on the platform is typically informal. Authors frequently ignore conventional grammar and punctuation rules.

Prior work on Turkish sarcasm detection has largely relied on instances typically drawn from Twitter or other microblog-style platforms and annotated for irony at the post level (Taslioglu and Karagoz, 2017; Ozturk et al., 2021; Dülger, 2018). More recently, context has been explored in Turkish news data by expanding the local sentence window within a paragraph to detect sarcasm (Eser and Bilgin, 2025). In this study, we introduce a Turkish sarcasm detection dataset from Ekşi Sözlük that provides title-level contextual summaries and enables comparisons between entry-only and context-aware modelling setups. We evaluate both entry-only and context-aware modelling setups to evaluate the contribution of contextual information to sarcasm detection. In this setting, the best overall

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 61–71

performance is obtained by fine-tuning BERTurk with title-level context, achieving 0.76 accuracy with balanced class-wise F1-scores (F1=0.77 for sarcasm, F1=0.75 for no-sarcasm). These results indicate that incorporating shared title context can improve sarcasm detection in multi-author discussion threads.

2 Related Work

Sarcasm detection is essential for many NLP tasks (Weitzel et al., 2016) and beyond. From a human-computer interaction perspective, humorous chatbots may increase user satisfaction (Shin et al., 2023). However, it is also important for systems to know when to use sarcasm and when to stick to literal explanations (Oprea et al., 2022). In this section, we give an overview of the sarcasm detection research in general before focusing on Turkish. We follow with the context-aware sarcasm detection datasets and review summarization literature that we leverage for context generation.

2.1 Sarcasm Detection

Sarcasm detection is studied as a text classification problem over short texts such as tweets and headlines. While English is the primary focus of much prior work, sarcasm detection is also explored in other world languages. Gong et al. (2020) introduces a large-scale Chinese sarcasm dataset, and Abu Farha and Magdy (2020) presents an Arabic Twitter corpus annotated for sarcasm. In addition, sarcasm detection is studied in multi-language settings such as Arabic–English (Abu Farha et al., 2022) and Czech–English (Ptáček et al., 2014).

In English, early work focused on Twitter and microblog texts (Barbieri et al., 2014; Ptáček et al., 2014). Subsequently, studies also consider news headlines (Shrikhande et al., 2020) and crowdsources corpora (Oraby et al., 2016). More recently, Oprea and Magdy (2020a) introduces an intended sarcasm dataset, where the annotators are the content creators themselves rather than the external annotators.

2.2 Turkish Sarcasm Datasets and Detection

It is not easy to precisely define sarcasm, where some sources take sarcasm as a subclass of irony (Leggitt and Gibbs, 2000), and some count them as separate notions (Ling and Klinger, 2016). Prior work on Turkish speakers also suggests that sarcasm is frequently used in Turkish, but the term itself has no exact equivalent; 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). Hence, prior datasets in Turkish have been published under varying task names (e.g., irony/satire), sometimes with overlapping conceptual scope. In this section, we summarize existing Turkish datasets.

Early Turkish microblog irony datasets are relatively small: Taslioglu and Karagoz (2017) samples a sentiment-stratified subset from Twitter, Dülger (2018) compiles a small mixed-source corpus (Twitter/microblogs), Karabaş and Dırı (2020) collects and normalizes tweets via the Twitter API. Later, Ozturk et al. (2021) introduces *IronyTR*, which is a balanced binary dataset collected from Twitter and other microblog platforms.

In the news domain, Onan and Toçoğlu (2020) constructs a large-scale Turkish satire corpus by collecting satirical articles from Zaytung and non-satirical news from the official Twitter page of media organizations. Most recently, Eser and Bilgin (2025) introduce a Turkish news-column dataset annotated into three classes (irony, sarcasm, normal) and explicitly investigate the role of context by constructing multiple dataset variants with increasing context width (from the target sentence alone to several preceding sentences within the same paragraph). In their setup, context is defined as a local sentence window within a single-author paragraph. However, in our setting, texts are Ekşi Sözlük entries written by multiple authors under a shared title, and the relevant context is the title-level discussion itself. (i.e., what the title is about and which shared topic is being referenced.) Accordingly, we represent context at the title level.

2.3 Context-Aware Sarcasm Detection

Beyond Turkish-only datasets, prior work has also examined sarcasm detection with contextual information. In this line of work, *context* is defined in different ways depending on the data source and the interaction setting.

Firstly, there are research that define *context* at the conversation level, typically as the dialogue history (i.e., one or more preceding turns in a thread) that the target utterance responds to (Ghosh et al., 2017, 2018; Ducret et al., 2020; Kim et al., 2024b; Srivastava et al., 2020). On the other hand, Oprea and Magdy (2019) defines *author context* as signals derived from a user's historical posts. Finally,

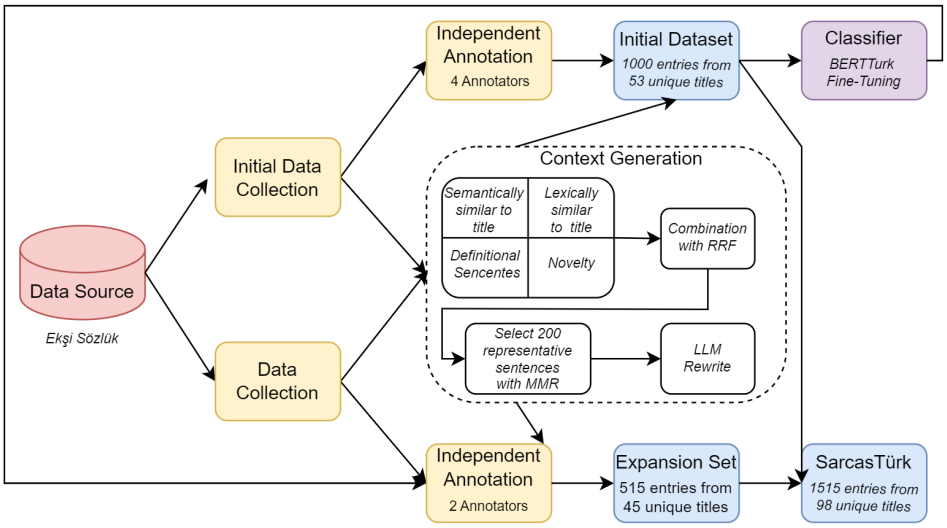

Figure 1: Overview of the dataset construction and context generation pipeline.

Khodak et al. incorporates *topic/thread level context*, where context provides access to the broader discussion structure (Khodak et al., 2018). Our work can also be situated within *topic-level context* modeling.

2.4 Summarization

Conversation and speech summarization have been widely studied in the literature; nevertheless, summarizing multi-speaker or discussion style content such as forums remains challenging (Fabbri et al., 2021). In this section, we give a review on forum and thread-like summarization discussions.

ForumSum (Khalman et al., 2021) is a large-scale conversation summarization dataset collected from diverse internet forums with human written abstractive summaries, enabling systematic study of multi speaker discussion summarization. *MREDDITSUM* (Overbay et al., 2023) is a multimodal abstractive summarization dataset of Reddit threads, where each instance includes the full discussion and associated images, supporting models that summarize grounded in both textual and visual cues. Beyond dataset construction, *MRCSum* (Kim et al., 2024a) propose a title-conditioned extractive approach that uses the document title as a query signal to select summary sentences.

In our setting, we similarly treat the Ekşi Sözlük title as a query signal, but the entries under that title written by different independent authors. Therefore, we leverage title conditioned sentence selection and redundancy control to produce a concise context representation at the title level.

3 SarcasTürk

In this section, we present SarcasTürk: a context-aware Turkish sarcasm detection dataset from Ekşi Sözlük, comprising two phases: initial dataset collection and expansion as shown in Figure 1.We first introduce the initial collection and how we generate title-level context to support context-aware experiments. Next, we describe a second data collection and annotation phase to expand the dataset.

3.1 Initial Data Collection & Annotation

The initial dataset was constructed manually by four native Turkish speakers familiar with Ekşi Sözlük's discourse norms. Every researcher browsed various titles across different domains and sentences from entries that appeared likely to contain sarcasm or intentional tone. Titles were chosen to represent a variety of domains including politics, popular culture, daily life, and humor to ensure diversity in tone and topic. In some cases, sarcasm was expressed only in a specific sentence or phrase within a longer entry. Instead of discarding these entries, the annotators extracted only the relevant portion containing the sarcastic expression. Because Ekşi Sözlük entries vary widely in length, tone, and narrative structure, this manual selection process allowed the dataset to capture naturally occurring examples of sarcasm as it appears in informal Turkish online discourse.

From each title, the researchers collected an approximately balanced set of sarcastic and non-sarcastic entries between October and December 2024, resulting in a total of 1000 entries from 53

Title	Context	Entry	Label
<Blinded> belediyesinin çocuklara sürprizi *<Blinded> municipality's surprise for children*	<Blinded> Belediyesi, çocukların oyun alanı taleplerine yanıt olarak okul bahçesine iki bank yerleştirmiştir. <cont'd>. *<Blinded> Municipality responded to children's requests for a playground by placing two benches in the schoolyard. <cont'd>*	o çocukların yerinde olmak vardı... *I wish I could be in those children's shoes...*	1 (Sarcasm)
<Blinded> belediyesinin çocuklara sürprizi *<Blinded> municipality's surprise for children*	<Blinded> Belediyesi, çocukların oyun alanı taleplerine yanıt olarak okul bahçesine iki bank yerleştirmiştir. <cont'd>. *<Blinded> Municipality responded to children's requests for a playground by placing two benches in the schoolyard. <cont'd>*	lan böyle bir saçmalığın yapılması kadar bunu kayda alıp iyi bi şeymiş gibi paylaşmak da ayrı bi şey. *doing something this ridiculous is one thing, but recording it and sharing it as if it's a good thing is something else entirely.*	0 (No Sarcasm)
anne babasıyla sigara içen 11 yaşındaki çocuk *an 11-year-old child smoking with their parents*	<Blinded>'te kaydedilen bir videoda, 11-12 yaşlarındaki bir çocuğun anne ve babasıyla birlikte sigara içtiği görülüyor. <cont'd>. *A video recorded in <Blinded> shows an 11-12-year-old child smoking with their parents. <cont'd>.*	çocuğuyla arkadaş gibi olabilen anne babalara hep imrenirim. * tebrik ederim. *I've always admired parents who can be like friends with their children. Congratulations.*	1 (Sarcasm)
anne babasıyla sigara içen 11 yaşındaki çocuk *an 11-year-old child smoking with their parents*	<Blinded>'te kaydedilen bir videoda, 11-12 yaşlarındaki bir çocuğun anne ve babasıyla birlikte sigara içtiği görülüyor. <cont'd>. *A video recorded in <Blinded> shows an 11-12-year-old child smoking with their parents. <cont'd>.*	çocuğun devlet tarafından alınmasını gerektirir. umarım görüntüler ilgili kişilere ulaşır. *This warrants the child being taken into state custody. I hope the footage reaches the relevant authorities.*	0 (No Sarcasm)

Table 1: Dataset examples with Turkish originals and English translations.

titles. Four researchers independently annotated all sarcastic entries with four different labels, depending on the clarity of the sarcasm level with scores: "Easy" (1), "Moderate" (2), "Hard" (3), and "No Sarcasm" (4). These four researchers then discussed entries in which the four labels varied greatly to agree on definitions and resolve conflicts. After this, the scores for all sarcastic entries were summed to reach a final label. The entries with scores between 4-6 received the label "Easy", 7-9 as "Moderate", 10-12 as "Hard", and 13-16 as "No Sarcasm". In the end, the dataset had 434 "Easy", 47 "Moderate", 16 "Hard", and 503 "No Sarcasm" entries. Three of the entries that were collected as sarcastic were labeled as "No Sarcasm" after the labeling process. While we have sarcasm-level granularity, we use only binary labels ("Sarcasm" and "No Sarcasm") for further analysis.

We labeled only the sarcastic entries and most of these were labeled "Easy" by the annotators. Hence, the annotated data is highly skewed. The kappa statistic is known to yield low values when the data are skewed (Viera et al., 2005). Hence, we do not report inter-annotator agreement using kappa scores. Before the discussion, 443 of 500 entries were unanimously labeled as sarcasm, and 490 of 500 were majority-voted as sarcasm. 402 out of 500 were majority-voted as "Easy".

3.2 Context Generation

In the nature of Ekşi Sözlük, a title page does not provide explicit conversational reply chains. Although entries under the same title are conceptually related, they do not form a clean dialogue structure. Also, an entry that appears purely normal in no-context may express a clear sarcasm once it's supported with the title-level context, as shown in the first row of Table 1. Therefore, rather than retrieving context on an utterance basis, we construct a shared contextual summary for each title. For this aim, we scraped all available entries for each title between September and October 2025 and used them to form a contextual summary that reflects the general theme. To do so, we created our own scraper that complies with Ekşi Sözlük's terms of service.

We construct four ranked sentence lists for each title: (i) For the *semantic* list, we embed all candidate sentences using the Sentence-Transformers model[1]. and rank them by cosine similarity to the title embedding. (ii) For the *lexical* list, we compute TF-IDF similarity between the title text and candidate sentences. (iii) Ekşi Sözlük's dictionary-like nature means that many titles contain definitional or descriptive sentences that serve as a neutral baseline for the discussion; to capture these, we build a *definitional* list using simple Turkish definitional cues and patterns, such as suffixes (e.g., *-dır/-dir/-dur/-dür*) that frequently appear in explanatory statements. (iv) Finally, we derive a *novelty* list by clustering sentence embeddings and selecting representative sentences from the clusters.

We combine the four ranked lists using Reciprocal Rank Fusion (RRF) to produce a single, unified ranking of candidate sentences. Next, we apply Maximal Marginal Relevance (MMR) with the goal of selecting sentences that remain highly relevant to the title while also avoiding near-duplicates among the selected items. 200 sentences per title were selected using this methodology.

Finally, after obtaining these MMR sets, we prompted multiple GPT models (GPT-o3-mini, GPT-4o, and GPT-4.1) to rewrite the 200 sentences into coherent title-level context paragraphs of 40-70 words. To validate the LLM-generated contexts, they were cross-checked by researchers who are familiar with the platform and the specific context of each title. While titles with inaccurate summaries were excluded, those that met our validation criteria were added to the dataset without any manual post-editing. With this procedure, we had three alternative context versions per title and therefore three context-aware dataset variants. We then fine-tuned the BERTurk classifier on each variant. Because GPT-4.1 produced slightly better F1 scores on the validation set, we use context generated by GPT-4.1 in the released dataset and in subsequent experiments.

3.3 Dataset Expansion

After the initial dataset was collected, we conducted a second data-collection phase in November 2025 that built on it. This expansion had two goals:(i) increasing the overall number of titles and entries, (ii) reducing the likelihood that the model memorizes misleading surface patterns. (e.g., *bknz.* "see also" was initially correlated with sarcastic predictions, so we added non-sarcastic *bknz.* examples during expansion).

3.3.1 Additional Title Scraping

One of the researchers, a native Turkish speaker and frequent user of Ekşi Sözlük, monitored the stream of titles in November 2025 and selected those that appeared likely to attract sarcastic contributions. Titles that were actively discussed around a specific news item, public incident, or viral phrase were prioritised, as these tend to generate a higher density of humorous and sarcastic entries over a distinct context. Typical examples include titles about a viral news item, a widely discussed public topic, or a phrase that became a meme.

After selecting the titles, we collected all associated entries with them and generated their contexts as detailed in Section 3.2. All collected entries from these newly sampled titles were first passed through the sarcasm classifier trained on the initial dataset. The model produced a binary prediction (Sarcasm vs. No Sarcasm) for every entry.

3.3.2 Filtering & Annotation

After the data collection, two human annotators independently assigned binary sarcasm labels to the entries considering the entry text and the context generated as specified in Section 3.2. In this annotation phase, researchers applied a set of filtering criteria as detailed below. Entries that failed any of these criteria were not annotated and were therefore excluded from the dataset. Only the entries that both annotators agreed on the label were

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

included in the dataset. In total, the dataset was expanded by 45 titles and 515 entries, yielding 98 unique titles and 1,515 entries overall. Eventually we have 774 Sarcasm and 741 No Sarcasm entries in SarcasTürk.

Filtering Criteria

Entries that exceeded a predefined length threshold, repeated essentially the same joke or sarcastic template as an already selected entry, or were only weakly related to the title or its overall topic were not included. We also excluded entries whose sarcastic meaning depended completely on out of context references to external events, people, or private conversations that could not be reconstructed from the title or context generated. In addition, entries whose main function was direct abuse or swearing, entries that were unintelligible due to severe spelling or grammatical issues, entries written predominantly in a foreign language or in highly atypical language, and entries that mainly redirected the reader to external resources (such as picture and video links or other titles and resources) rather than contributing their own content were discarded. When integrating the newly labelled entries into the corpus, we preserved the overall class balance between sarcastic and non-sarcastic texts. In cases where one class was over-represented within a title or in the expansion batch as a whole, surplus entries from the majority class were discarded.

4 Baseline Experiments

As a baseline, we frame sarcasm detection as a binary sentence-level classification task. Each entry from Ekşi Sözlük is assigned one of two labels, *1 (Sarcasm)* or *0 (No Sarcasm)*, based on the manual annotation procedure described in Section 3. We use three splits: a training set for fine-tuning, a validation set, and a test set for reporting final results in SarcasTürk. Encoder-based models are fine-tuned on the training split and selected using the validation split, whereas decoder-only LLM baselines are evaluated in a zero-shot setting without any fine-tuning. In all experiments, we report overall accuracy and class-wise F1-scores for the sarcastic class (F1(1)) and the non-sarcastic class (F1(0)).

For encoder-based models, we fine-tune BERTurk [2] (Schweter, 2020) as a sentence-level classifier. In the **Entry-Only** configuration, only the entry text is given to the model. We use a standard classification head on top of the [CLS] representation and fine-tune the model for five epochs with a batch size of 16, maximum sequence length 384 and learning rate 4×10^{-5}. In the **Context-Aware** configuration, we use the title-level contexts introduced in Section 3.2. Entry and context are encoded separately using a shared BERTurk encoder with fixed token budgets of 96 tokens for the entry and 288 tokens for the context. This dual-encoder model is trained with the same optimization scheme as the entry-only model.

For decoder-only large language models, we do not fine-tune the models on our dataset. Instead, we treat them as zero-shot sarcasm classifiers accessed via their chat-completion APIs. We consider three models: GPT-4o, GPT-4.1, and Llama-3.3-70B. All three are prompted with the instruction that defines the task and constrains the output to a `label`. In the **No Context** condition, the user message contains only the entry text. In the **Context-Aware** condition, we add the context. All zero-shot LLMs are evaluated on the same test set as the fine-tuned BERTurk models.

5 Results & Discussion

Table 2 summarizes the performance of all models with and without title-level context. In the table, we show the model performance on the initial dataset and SarcasTürk. Overall, the best results are obtained by the fine-tuned BERTurk model in the context-aware configuration on SarcasTürk, which reaches an accuracy of 0.76 with balanced F1-scores for both classes (F1(1) = 0.77, F1(0) = 0.75). The entry-only BERTurk baseline already performs strongly (0.73 accuracy), but adding context yields consistent gains in both F1(1) and F1(0), suggesting that the title-level summaries provide useful cues for disambiguating sarcastic and non-sarcastic entries. A similar effect holds on the initial dataset: BERTurk achieves the strongest non-sarcastic performance (F1(0) = 0.76) in both configurations, while context primarily boosts sarcasm recognition (F1(1) from 0.66 to 0.72).

Among the zero-shot LLMs, GPT-4.1 achieves the strongest overall performance. Without context, GPT-4.1 reaches 0.73 accuracy and F1(1) = 0.76, closely matching the entry-only BERTurk baseline and outperforming it on the sarcastic class while lagging behind on the non-sarcastic class (F1(0) = 0.68 vs. 0.74). On the initial dataset, GPT-

[2] https://huggingface.co/dbmdz/bert-base-turkish-uncased

Model	Initial Dataset			SarcasTürk		
	Accuracy	F1(1)	F1(0)	Accuracy	F1(1)	F1(0)
Llama-3.3-70B Zero Shot (No Context)	0.63	0.67	0.58	0.62	0.67	0.56
Llama-3.3-70B Zero Shot (Context-Aware)	0.66	0.72	0.58	0.65	0.71	0.55
GPT-4o Zero Shot (No Context)	0.63	0.69	0.52	0.63	0.70	0.50
GPT-4o Zero Shot (Context-Aware)	0.66	0.74	0.51	0.67	0.74	0.53
GPT-4.1 Zero Shot (No Context)	0.74	0.77	0.71	0.73	0.76	0.68
GPT-4.1 Zero Shot (Context-Aware)	0.70	0.75	0.62	0.71	0.76	0.63
BERTurk (No Context)	0.72	0.66	0.76	0.73	0.71	0.74
BERTurk (Context-Aware)	0.74	0.72	0.76	**0.76**	**0.77**	**0.75**

Table 2: Model performance on initial dataset and SarcasTürk (1 = Sarcasm, 0 = No Sarcasm).

4.1 shows the same pattern and slightly higher scores in the no-context setting (0.74 accuracy; F1(1) = 0.77, F1(0) = 0.71). Adding context to GPT-4.1 leads to a small drop in accuracy (0.71) and in F1(0), while leaving F1(1) unchanged. This drop is also observed on the initial dataset (0.70 accuracy; F1(0) from 0.71 to 0.62). A qualitative analysis of the errors for GPT-4.1 reveals that this drop in F1(0) is largely driven by a tonal mismatch. Since the LLM-generated contexts have a typically objective and encyclopedic tone, the model tends to misinterpret the sharp contrast between this formal background and the informal, subjective nature of user entries as a sign of irony. Consequently, for this model, in some cases, context acts as noise that triggers false positives in the non-sarcastic class.

GPT-4o and Llama-3.3-70B show a clearer benefit from context. For GPT-4o, context increases accuracy from 0.63 to 0.67 and improves F1(1) from 0.70 to 0.74 and F1(0) from 0.50 to 0.53. The same situation holds on the initial dataset, where context raises GPT-4o from 0.63 to 0.66 accuracy and improves F1(1) from 0.69 to 0.74. Llama-3.3-70B exhibits a similar pattern on the sarcastic class: F1(1) rises from 0.67 to 0.71, and accuracy improves from 0.62 to 0.65 when context is provided. On the initial dataset, context improves Llama-3.3-70B from 0.63 to 0.66 accuracy and from 0.67 to 0.72 in F1(1). However, the gains in sarcasm detection come at the cost of a slight decrease in F1(0), indicating that these models tend to predict *Sarcasm* more aggressively when context information is available.

Across all zero-shot LLMs, models are competitive at recognising sarcastic entries but are less reliable at correctly identifying non-sarcastic text, where they often produce false positives. In contrast, fine-tuned BERTurk, especially in the context-aware setting, maintains a better balance between the two classes.

5.1 Limitations and Future Work

SarcasTürk was created in a two-stage setup; the initial step and the expansion phase. In the initial step, we collected and annotated the data without context. We then created a context-generation workflow and augmented our initial dataset with title-level context.

Since the initial dataset was labelled without context, we didn't collect any entries from the titles that may require context to understand sarcasm (e.g., as shown in the first row of Table 1 where we need context to decide on sarcasm). In the dataset expansion part, we chose entries without this limitation. Since we have a different filtering methodology, we show the results for the initial dataset separately in Table 2.

We extracted only the sarcastic parts of some of the long entries during the initial data collection step. However, considering the full text of the entries could also be helpful, as it provides context. Since the initial data collection was in Fall 2024, some of these entries are inaccessible (e.g., either deleted by the users or the platform). Hence, we cannot reach the full text for such entries. Future work could collect entries as a whole text to make further analysis.

Another limitation arises from the generated context itself. As observed with GPT-4.1 and discussed in Section 5, the formal tone of LLM-generated contexts can create a tonal mismatch with informal entries. This causes false positives in the non-sarcastic class. Finally, while we intentionally balanced the dataset to ensure sufficient supervision

for the sarcastic class, real-world sarcasm distribution is typically highly skewed. Future evaluations on unbalanced subsets would be valuable to assess model robustness in realistic scenarios.

We release SarcasTürk, a Turkish sarcasm detection dataset with title-level context, and provide baseline results with BERTurk to facilitate future research. Future works may (i) expand the dataset using the proposed expansion workflow to generalize more domains, (ii) explore alternative forms of context construction and assess how context affects the performance, (iii) position SarcasTürk within the multilingual sarcasm literature by preparing a standardized evaluation protocol, and (iv) conduct more extensive benchmarking with other state-of-the-art models that are open-sourced, as well as other encoder-only models such as mmBERT (Marone et al., 2025).

6 Ethical Considerations

This study aims to support research on sarcasm detection and context-aware language understanding in Turkish online discourse. SarcasTürk is constructed from naturally occurring entries on Ekşi Sözlük and is intended solely as a resource for academic research on sarcasm detection and related language technologies, not for amplifying offensive content or targeting individuals or groups.

Because the data is collected from a large, user-generated platform and reflects naturally occurring language use in Turkish online discourse, it contains sensitive and potentially harmful material that is commonly present in such environments. Entries may include racist or sexist remarks, coarse slang, swear words, sexual content, and mocking or demeaning comments about topics, social groups, or well-known public figures. Many entries express strong opinions, generalisations, or exaggerations and may explicitly or implicitly single out public figures or communities. We did not censor such content to preserve the original content. Readers should therefore be aware that the dataset contains offensive and disturbing language.

All entries in the corpus were selected and annotated by native Turkish speakers who are familiar with Ekşi Sözlük and with the goals of the project. Before the data collection, the annotators were agreed on the nature of the public content, and those who were uncomfortable with such material did not participate. During annotation and data collection, researchers focused only on assessing the presence or absence of sarcasm. They did not rate entries according to their moral acceptability, political stance, or factual correctness.

The dataset contains only four columns: title, context, entry, label. We don't release usernames or any other metadata that could be used to identify the Ekşi Sözlük authors.

Despite these precautions, the dataset may still reflect social biases present in the source platform. Certain groups, topics, or styles of expression may be over or under-represented, and models trained on this data can inherit or amplify such biases, particularly towards communities or polarising political topics.

We used Grammarly and ChatGPT 5.1 to check spelling and text flow after drafting our own version, and incorporated suggestions for more appropriate wording.

7 Conclusions

In this work, we introduced SarcasTürk, a context-aware Turkish sarcasm detection dataset collected from Ekşi Sözlük. SarcasTürk contains 1,515 entries from 98 titles with binary sarcasm labels and title-level context. We also present a context generation pipeline that constructs these title-level contexts, which enables context-aware modeling and evaluation. Our baseline model shows that context helps sarcasm detection: context-aware BERTTurk achieves the best overall performance (0.76 accuracy with balanced class-wise F1 scores).

Acknowledgements

This work was supported, in part, by Sabanci University, project number B.A.CF-25-03063.

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 Branowska, 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.

Maud Ducret, Ludwig Kruse, Carla Martinez, Anna Feldman, and Jing Peng. 2020. You don't say... linguistic features in sarcasm detection. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, Torino, Italy. Accademia University Press.

Oğuzhan Dülger. 2018. Türkçe metinlerde roni tespiti (Irony Classification in Turkish Text). In *Proceedings of the 12th Turkish National Software Engineering Symposium (UYMS 2018)*, volume 2201 of *CEUR Workshop Proceedings*, page –. CEUR-WS.org.

Ekşi Sözlük. 1999. Ekşi sözlük.

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).

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Debanjan Ghosh, Alexander R. Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany. Association for Computational Linguistics.

Raymond W Gibbs. 2000. Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.

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.

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.

Misha Khalman, Yao Zhao, and Mohammad Saleh. 2021. ForumSum: A multi-speaker conversation summarization dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4592–4599, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024a. Title-based extractive summarization via MRC framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16175–16186, Torino, Italia. ELRA and ICCL.

Yumin Kim, Heejae Suh, Mingi Kim, Dongyeon Won, and Hwanhee Lee. 2024b. KoCoSa: Korean context-aware sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9890–9904, Torino, Italia. ELRA and ICCL.

John S Leggitt 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.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Diana G Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Lrec 2014 proceedings*. ELRA.

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.

Silviu Oprea and Walid Magdy. 2019. Exploring author context for detecting intended vs perceived sarcasm.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2859, Florence, Italy. Association for Computational Linguistics.

Silviu Oprea and Walid Magdy. 2020a. 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.

Silviu Vlad Oprea and Walid Magdy. 2020b. The effect of sociocultural variables on sarcasm communication online. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–22.

Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. Should a chatbot be sarcastic? understanding user preferences towards sarcasm generation. In *60th Annual Meeting of the Association for Computational Linguistics*, pages 7686–7700. 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.

Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. 2023. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132, Singapore. 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.

Penny Pexman, Lorraine Reggin, and Kate Lee. 2019. Addressing the challenge of verbal irony: Getting serious about sarcasm training. *Languages*, 4(2):23.

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.

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

Hyunju Shin, Isabella Bunosso, and Lindsay R Levine. 2023. The influence of chatbot humour on consumer evaluations of services. *International Journal of Consumer Studies*, 47(2):545–562.

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.

Himani Srivastava, Vaibhav Varshney, Surabhi Kumari, and Saurabh Srivastava. 2020. A novel hierarchical BERT architecture for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 93–97, Online. Association for Computational Linguistics.

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.

Anthony J Viera, Joanne M Garrett, and 1 others. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland. Association for Computational Linguistics.

Leila Weitzel, Ronaldo Cristiano Prati, and Raul Freire Aguiar. 2016. The comprehension of figurative language: What is the influence of irony and sarcasm on nlp techniques? In *Sentiment analysis and ontology engineering: An environment of computational intelligence*, pages 49–74. Springer.

A LLM System Message for Context Generation

```
- Rolün: Başlık altındaki aday cümleleri yalnızca
↪  verilen içerikten hareketle nötr,
↪  bilgi-odaklı bir bağlama sıkıştıran editör.
- Amaç: Aşağıdaki 200 cümleden yararlanarak,
↪  {TITLE} başlığının bağlamını açıklayan 3-4
↪  cümlelik, okunur ve kapsayıcı bir paragraf
↪  üret.
- Girdi:
-- Başlık: {TITLE}
-- Aday cümleler (her biri bir satır):
- Kurallar:
-- Sadece verilen cümlelerdeki bilgiye dayan;
↪  harici bilgi ekleme/tahmin yapma.
-- Nötr/ansiklopedik ton: 1./2. tekil/çoğul kişi
↪  (ben, biz, sen, siz) ve duygusal/argo
↪  kullanma.
-- Tanıma yakın giriş + tekrarlanan temalar (örn.
↪  özellikler, tartışma eksenleri, tipik
↪  örnekler) + varsa mizah/sarkazm üslubuna üst
↪  düzey atıf.
-- Liste-ezberi (X, Y, Z gibi) ve marka/model
↪  şakalarını öz haline getir; isimleri yığma.
-- Çelişki varsa üst seviye birleştir: "bazı
↪  kullanıcılar . . . , diğerleri . . . ".
-- Başlığı papağanlama (sadece "{{TITLE}} . . . "
↪  demek) yapma; içerik taşı.
```

```
-- Uzunluk: 3-4 cümle, toplam 40-70 kelime
↪  civarı.
- İş akışı:
1) Cümleleri hızlıca tara → tekrar eden temaları
↪  ve tanımsal ipuçlarını bul.
2) Aşırı öznel/argo/kişisel anı cümlelerini
↪  özetleyerek nötrleştir.
3) Bir paragraf yaz: (i) konu/alan çerçevesi,
↪  (ii) ana temalar, (iii) varsa karşıt
↪  görüş/mizah notu.
- Çıktı (yalnızca paragraf):
-- Türkçe, tek paragraf, 3-4 cümle.
-- Başlık veya köşeli parantez/ID yazma; sadece
↪  özet paragrafını ver.
- Veri:
-- Başlık: {{TITLE}}
-- Cümleler: {{SENTENCE_BLOCK_OF_200}}
```

In our context generation step, the LLM was instructed using the Turkish system message. We include the English translation below for clarity.

```
- Your role: An editor who compresses the
↪  candidate sentences under the title into a
↪  neutral, information-focused context,
↪  relying only on the provided content.
- Goal: Using the 200 sentences below, produce a
↪  readable and comprehensive paragraph of 3–4
↪  sentences that explains the context of the
↪  {TITLE}.
- Input:
-- Title: {TITLE}
-- Candidate sentences (each on its own line):
- Rules:
-- Rely only on the information in the given
↪  sentences; do not add external information or
↪  make guesses.
-- Neutral/encyclopedic tone: do not use 1st/2nd
↪  person singular/plural (I, we, you) and avoid
↪  emotional language or slang.
-- A definition-like opening + recurring themes
↪  (e.g., characteristics, axes of debate,
↪  typical examples) + if present, a high-level
↪  reference to humor/sarcasm style.
-- Condense list-like recitations (e.g., X, Y, Z)
↪  and brand/model jokes into their essence; do
↪  not pile up names.
-- If there are contradictions, merge at a high
↪  level: “some users . . . , others . . . ”.
-- Do not parrot the title (i.e., avoid merely
↪  saying “{{TITLE}} . . . ”); convey substance.
-- Length: 3–4 sentences, around 40–70 words
↪  total.
- Workflow:
1) Skim the sentences quickly → find recurring
↪  themes and definitional cues.
2) Neutralize overly
↪  subjective/slang/personal-anecdote sentences
↪  by summarizing them.
3) Write one paragraph: (i) topic/field framing,
↪  (ii) main themes, (iii) if applicable, a note
↪  on opposing views/humor.
- Output (paragraph only):
-- Turkish, single paragraph, 3–4 sentences.
-- Do not write the title or brackets/IDs;
↪  provide only the summary paragraph.
- Data:
-- Title: {{TITLE}}
-- Sentences: {{SENTENCE_BLOCK_OF_200}}
```

B Zero Shot LLM Test System Messages

B.1 No Context

```
You are a sarcasm classifier. The user will
↪  provide a piece of text
The goal is to decide if the Text itself is
↪  sarcastic or not.
Respond ONLY in the following JSON format WITHOUT
↪  code fences and without any extra text:
 '{ "label": "Sarcasm" | "No Sarcasm", "reason":
 ↪  "<short reason>" }\n'
If uncertain, choose the most likely label and
↪  state that in the reason.
```

B.2 Context-Aware

```
You are a sarcasm classifier. The user will
↪  provide a piece of text and context.
The goal is to decide if the Text itself is
↪  sarcastic or not.
The Context is only background information to
↪  help you interpret the Text
Respond ONLY in the following JSON format WITHOUT
↪  code fences and without any extra text:
'{ "label": "Sarcasm" | "No Sarcasm", "reason":
↪  "<short reason>" }\n'
If uncertain, choose the most likely label and
↪  state that in the reason.
```