Turkish Delights: A Dataset on Turkish Euphemisms

Hasan Can Biyik and **Patrick Lee** and **Anna Feldman**
Montclair State University
New Jersey, USA
{biyikh1,leep,feldmana}@montclair.edu

Abstract

Euphemisms are a form of figurative language that is relatively understudied in natural language processing. This research extends the current computational work on potentially euphemistic terms (PETs) to Turkish. We introduce the Turkish PET dataset, the first available of its kind in the field.[1]. By creating a list of euphemisms in Turkish, collecting example contexts, and annotating them, we provide both euphemistic and non-euphemistic examples of PETs in Turkish. We describe the dataset and methodologies and also experiment with transformer-based models on Turkish euphemism detection by using our dataset for binary classification. We compare performances across models using F1, accuracy, and precision as evaluation metrics.

1 Introduction

Euphemisms are polite or indirect words or expressions used in substitution of unpleasant or more offensive ones. They can be used to show kindness while discussing sensitive or taboo topics (Bakhriddionova, 2021) such as saying *between jobs* instead of *unemployed*, or as a way to make unpleasant or unappealing things sound less harsh (Karam, 2011), such as saying *passed away,* instead of *died*. Similar to the word *died* in English, Turkish makes use of many substitutions for the word *öl-mek/öl-dü (to die/died)*, which is considered unpleasant. The substitutions for this word could be given as *vefat etmek (to pass away)*, *öbür dünyaya göçmek (to migrate to the other world)*, *hakkın rahmetine kavuşmak (to go to kingdom come)*. Euphemisms can be used to conceal the truth (Rababah, 2014); for instance, if one were to use the expression *enhanced interrogation techniques*, one would mean *torture* (Lee et al., 2022b). Furthermore, humans may not agree on what a euphemism is (Gavidia et al., 2022a). There are various challenges regarding euphemisms. For instance, in some cases, words or expressions might develop or lose euphemistic meanings in time (Pinker, 1994, 2003). Due to the aforementioned reasons, the words and phrases in this research will be referred to as *potentially euphemistic terms (PETs)* (Lee et al., 2022c). Euphemisms pose a challenge to Natural Language Processing (NLP) due to this figurative behavior as they might also have a non-euphemistic interpretation in certain contexts. For example, while the Turkish PET *mercimeği fırına vermek* means *to put the lentil in the oven* literally, it could mean *to have sex/to get someone pregnant* euphemistically. In the following sentence, this PET used literally: "Günümüzde hem <mercimeği fırına vermek> daha kolay, hem de fırında makarna yemek..." which can be translated as "Nowadays, it's easier to <put the lentil in the oven> and to eat mac and cheese..." However, it was used euphemistically in the following sentence: "Gel gör ki kasabanın yegane doktoru ile pişiren bu kadın, zaman zaman <mercimeği fırına veriyorlarmış>" which can be translated as "However, it turns out that this woman, who is having an affair with the town's only doctor, sometimes <puts the lentil in the oven>" meaning that the doctor and woman are involved in a secretive or intimate sexual intercourse.

Conducting a euphemism detection task in Turkish has several challenges to overcome. Firstly, to the best of our knowledge, there are no available datasets for automatic euphemism detection task in Turkish. Academic research, published books, articles, and other resources on this topic are very limited, making the collection of PETs difficult. In this research, we aim to identify PETs in Turkish and create a dataset of Turkish PETs with the help of native Turkish annotators who have a linguistics background. We fine-tune language models (LMs) such as BERTurk (DBMDZ, 2019; Beyhan et al., 2022) and ELECTRA (Clark et al., 2020) and large

[1]The dataset is available at https://github.com/hasancanbiyik/Turkish_PETs

language models (LLMs), such as XLM-RoBERTa (AI, 2019; Conneau et al., 2020) and mBERT (AI, 2018; Devlin et al., 2019) for euphemism detection in Turkish. Therefore, the significant contributions of this paper are as follows:

- Introduction of the Turkish PETs dataset,
- Overview of the collection and annotation process of Turkish PETs,
- Comparison of the performances of XLM-RoBERTa, mBERT, BERTurk, and ELECTRA in detecting PETs in Turkish, using F1, accuracy, and precision as evaluation metrics,
- Cross-linguistic comparison of PETs in Turkish and other languages, accompanied by an analysis of potentially interesting patterns.

Additionally, through extending euphemism detection task to a new language, we contribute to a better understanding of how euphemisms are utilized and interpreted across different linguistic and cultural contexts.

2 Turkish Language

Agglutinative languages, such as Turkish, form words by adding multiple affixes to a stem, with each affix representing a distinct morphological feature (Comrie, 1988). This morphological productivity creates a vast number of possible word forms, making it difficult to develop comprehensive dictionaries or rule-based systems for tasks like euphemism detection. For instance, the PET *hayata gözlerini yummak* (*to close one's eyes to life*) can be formed as yum-***du***, yum-***muş***, yum-***duğunda***" and many other variations. See Table 1 for more examples regarding morphological variations.

The free word order in Turkish, where the position of words in a sentence can vary without significantly changing the meaning (Göksel and Kerslake, 2004), poses another challenge for euphemism detection. This flexibility makes it difficult to rely on fixed patterns or word sequences to identify euphemisms. For example, the PET *uyutmak* (*to put to sleep*) can appear in various positions within a sentence, making it harder to detect reliably.

Similar to euphemisms in other languages, the meaning of words and expressions are context dependent in Turkish. While one word can be used euphemistically in one sentence, it might not have euphemistic meaning in another. For instance, the PET *engelli* might be used euphemistically to indicate that the person is *disabled*, but it might also have its non-euphemistic meaning of *blocked*.

Moreover, Turkish is considered to be a low-resource language because of the limited availability of annotated datasets. It was also stated by various researchers that collecting data from various sources and labeling them was a challenging process (Mutlu and Özgür, 2022). Since there was no available dataset that contained euphemisms in Turkish with examples, it was necessary for us to build a dataset and get it annotated by native Turkish annotators.

3 Automatic Euphemism Detection

Euphemism detection can be viewed as a classification task in which an input text is classified as containing a euphemism or not.

While this can be theoretically done at at the phrase-level or sentence-level euphemism detection, previous work has focused on classifying examples containing specific multi-word expressions, which may or may not be used euphemistically depending on the context (Lee et al., 2022a). A number of approaches have performed decently at the task using language models such as transformers, improving upon baselines using various techniques. For example, Keh et al. (2022) use an ensemble of models each utilizing a combination of data and contextual augmentations to improve performance by 5 Macro-F1 points. Kesen et al. (2022) achieve similar improvements by incorporating non-euphemistic meanings and image embeddings associated with PETs. Maimaitituoheti et al. (2022) propose a prompt-based approach for euphemism detection utilizing the language model RoBERTa, achieving an F1 score of 85.2%, demonstrating the effectiveness of prompt-based learning. Similar to our initial dataset, which contained more than 6,000 examples, the dataset they used was imbalanced and had more euphemistic examples than non-euphemistic. They noted the model's superior performance on euphemistic sentences compared to non-euphemistic ones due to this imbalance.

Given the nuanced nature of these expressions in the Turkish language and the lack of previous work on figurative language processing in Turkish, this study aims to investigate how well different language models identify and categorize PETs in Turkish. We fine-tuned two large multilingual models, XLM-RoBERTa and mBERT, along with lan-

PET	Variations (Turkish)	Variations (English Equivalents)
aramızdan ayrıldı (left us)	aramızdan ayrıldı, aramızdan ayrılışının, aramızdan ayrılan, aramızdan ayrılanlar, aramızdan ayrılması, aramızdan ayrılalı	(has) left us, of his/her/their departure from us, the one who left us, those who left us, his/her/its departure from us, since he/she/they left us
beklemek (to expect)	bekliyor, bekliyoruz, bekleyen, bekledikleri, bekleniyor, bekleyeceğiz, bekliyorsunuz (...)	is expecting, we are expecting, the one who is expecting, what they are expecting for/whom they are expecting for/that they are expecting for, is being expected/is expected, we will expect, you are expecting (plural or formal)
hakka yürümek (to walk to God)	hakka yürüyen, hakka yürümesinden, hakka yürüdü, hakka yürümüştür	the one who walked to God, from his/her/their walking to God, walked to God, has walked to God

Table 1: Examples and morphological variations of Turkish PETs

guage models specifically trained on extensive corpora of Turkish text data: BERT-base-turkish-cased and ELECTRA-base-turkish-cased-discriminator. These models were chosen to examine the impact of model size, training data, and architecture on euphemism detection performance. We hypothesized that XLM-RoBERTa and mBERT would provide strong general language understanding capabilities, as large multilingual models are trained on vast amounts of diverse data. On the other hand, BERT-base-turkish-cased and ELECTRA-base-turkish-cased-discriminator, being specifically trained on Turkish text, were hypothesized to capture more nuanced aspects of euphemistic language in Turkish due to their exposure to a wider range of Turkish expressions and linguistic patterns.

Our focus on the Turkish language addresses a gap in existing research, as most previous studies have primarily concentrated on English euphemisms (Felt and Riloff, 2020; Zhu and Bhat, 2021; Zhu et al., 2021; Gavidia et al., 2022a,b; Lee et al., 2022a, 2023). By extending the euphemism detection task to a new language, we contribute to a better understanding of how euphemisms are utilized and interpreted across different linguistic and cultural contexts. The recent Multilingual Euphemism Detection Shared Task by Lee and Feldman (2024) has encouraged researchers to explore multilingual and cross-lingual methods for identifying euphemisms. This research emphasizes the importance of understanding euphemisms in different languages.

4 Data Collection and Annotation

4.1 Data Collection

To find PETs in Turkish, we analyzed the PETs in other languages described in previous work (Lee et al., 2023, 2024), such as American English, Mandarin Chinese, Yorùbá, and a mix of Spanish dialects to see whether there were overlapping words or expressions used euphemistically (see Table 2). As a result, we were able to compile an initial list of Turkish PETs.

Through reviewing published articles and papers related to euphemisms in Turkish, such as those by Aksan (1994); Karabulut and Ospanova (2013); Çabuk (2015), we expanded our list of PETs. Another method we used to collect PETs was by posting polls on social media. Initially, we explained the concept of "PETs" and provided examples. We then utilized social media to share these polls, where Turkish native speakers could share their ideas for new PETs. As a result, our Turkish PETs list now comprises a total of 122 entries. We also included detailed information for each PET, such as euphemistic category (e.g. bodily functions), meaning, non-euphemistic meaning, literal translation, and the source it was from. The list is categorized into 10 groups with varying frequencies, which can be seen in Table 4 . These categories were created based on the characteristics of the PETs. For example, the PET "*görme engelli*" (*visually impaired*) is related to physical attributes, and therefore it was added to the "physical/mental attributes" category.

Once the PETs list was finalized, we utilized a

English	Chinese	Spanish	Turkish	Yoruba
adult beverage	-	✓	✓	✓
birds and the bees	-	-	-	-
economical	✓	✓	✓	✓
pass away	✓	✓	✓	✓
pro-life	-	✓	-	-
under the weather	-	-	-	-

Table 2: Examples of (non-)overlapping PETs across the five languages.

Turkish corpus known as the TS Corpus Project (Sezer, 2017). We selected TS Corpus v2 and TS Timeline Corpus. TS Corpus v2 drew from the BOUN Web Corpus and included 491,360,398 tokens and 4,950,407 word types. TS Timeline Corpus contained more than 700 million tokens and over 2.2 million news and articles. To search for texts containing PETs for binary classification purposes, we utilized regular expressions, accounting for the agglutinative nature of the Turkish language. This approach allowed us to capture various word forms effectively. For instance, for the PET *hamileliği sonlandırmak* (*to terminate pregnancy*), we designed a regular expression to detect all variations of hamile-*lik* (pregnancy), hamile-*liğini* (her pregnancy), hamile-*liğimi* (my pregnancy), sonlan-*dırdı* (terminated/has terminated), sonlan-*dıracakmış* (I heard that she will terminate), sonlan-*dıramadı* (she could not terminate), etc., r"(hamileli\w+ sonlan\w+)". As a result, we successfully captured variations of each PET were successfully captured. These captured PETs were extracted and highlighted within their sentences using brackets, as shown: "Duyduğuma göre arkadaşı <hamileliğini sonlandırmış>." (I heard that her/his friend will <terminate her pregnancy>.) Additionally, we included preceding and succeeding sentence(s), if available, to form the entire example context for that PET. These contexts usually consisted of four sentences at most. Not all PETs on the initial list were found in the corpus; of the 122, only 58 were found and have at least one example. These examples were then compiled for the annotation phase.

4.2 Annotation

Annotators were provided text examples (∼1-4 sentences) of PETs in context, as can be seen in Table 3. To recruit Turkish annotators, we utilized social media platforms to find volunteers with a background in linguistics or an interest in the field. After several informational meetings, the annotators were briefed about the research purpose, the annotation process, and the concept of PETs. These meetings were recorded with the consent of the annotators. They were instructed to label the examples as "1" if the highlighted word or expression was used euphemistically, and as "0" if it was not. Following the completion of all annotations, an additional meeting was held to address any disagreements. During this discussion, some labels were revised. Notably, examples that received conflicting labels from the annotators—euphemistic by two and non-euphemistic by another two—had to be excluded from the dataset. This underscored the inherent challenges humans face in consistently interpreting whether a word or expression is used euphemistically.

For the annotation task, we divided the volunteers into five groups, with each group comprising three annotators. The first group annotated 975 examples, the second group annotated 1200 examples, the third group annotated 1300 examples, the fourth group annotated 1099 examples, and the fifth group annotated 1500 examples. As a result, there were 6,074 annotated examples at the end of the annotation task. Subsequently, each group's examples were annotated by one annotator from another group—for instance, an annotator from the first group annotated the second group's examples, and so on, ensuring each example was annotated by four different people. Throughout this process, examples with discrepancies were highlighted for further discussion during a recorded meeting with the available annotators. Disagreements were resolved by majority vote to finalize the labels. However, examples receiving split decisions (two annotators labeling euphemistic and two labeling non-euphemistic) were removed from the dataset. Sample examples and their final annotated labels can be found in Table 3.

While each example ultimately had four sepa-

PET	Label	Example
uyutmak	*euphemistic*	(...) Hollywood'un en çok tanınan köpekleri arasında yer alan Jack Russell cinsi Uggie <uyutularak> yaşamına son verildi. Uggie, katıldığı Oscar gecesiyle ününe ün katmış ve Cannes'da Palm Dog Ödülü'nün de bulunduğu birçok ödül kazanmıştı. (...) / One of Hollywood's most well-known dogs, the Jack Russell Terrier named Uggie, was <put to sleep>. Uggie gained even more fame by attending the Oscars and won many awards, including the Palm Dog Award at Cannes.
	non-euphemistic	İNSANA en çok benzeyen hayvan olarak bilinen şempanzeler, yavrularını titizlikle büyütüyor. Anne şempanze, yavrusunu kucağında <uyutuyor> ve gerektiğinde battaniyeyle üstünü örtüyor. (...) / Chimpanzees, known as the animals most similar to humans, meticulously raise their young. A mother chimpanzee <puts her baby to sleep> in her arms and covers it with a blanket when necessary.
muayyen günü	*euphemistic*	Kadınların <muayyen günleri> ya da hamilelik dönemlerinin de gözetilmesi amacıyla, nöbet ve görevlendirme sürelerine yeni esaslar getirilirken, muharebe eğitiminde el bombasını atma kurallarının bile kadınlar gözetilerek yeniden düzenlenmesi, Askerlik erkek işidir diyenleri dehşete düşürüyor." / In order to account for women's <specific days> or pregnancy periods, new principles have been introduced regarding the duration of duty and assignments. Even the rules for throwing grenades in combat training have been rearranged with women in mind, which horrifies those who say "military service is a man's job."
	non-euphemistic	Davetiyede, dispeç ile müsbit vesikaların mahkeme kaleminde incelenebileceği ve çağırılanın daha önce de dispeçe karşı mahkemede itirazda bulunabileceği <muayyen günde> gelmediği takdirde dispeçe muvafakat etmiş sayılacağı yazılır." / The invitation states that the dispatch and supporting documents can be reviewed in the court clerk's office, and that if the summoned party, who could have previously objected to the dispatch in court, does not appear on the <specified day>, they will be deemed to have consented to the dispatch.
ince hastalık	*euphemistic*	Eleni zamanında Eftelya'nın anneannesini yakalandığı <ince hastalık>tan Kerim hocanın iyileştirdiğini ve bunu da aileden gizli yaptığını anlatır. / Eleni explains that in the past, Kerim Hoca cured Eftelya's grandmother of <thin disease> and that he did this secretly, without the family's knowledge.
	non-euphemistic	Burdaki balların her derde deva olduğunu, <ince hastalık>lara iyi geldiğine inaüıñak'¸bu nedenle de ilaç olarak kullanılmaktadır. / The honey here is believed to be a cure for every ailment and is therefore used as medicine, particularly for treating <thin diseases>.

Table 3: Euphemistic and Non-euphemistic Usages of PETs

rate annotations, the annotators were allowed to collaborate and influence each others' opinions, nullifying potential inter-rater agreement analyses. We instead conducted inter-rater agreement analysis on a subset of 396 examples, labeled by two annotators who primarily worked separately. Co-

hen's kappa for these two raters was 0.696, which is rated as moderate to substantial agreement (Cohen, 1960). Interestingly, Krippendorf's alpha was 0.693, which is higher but still largely comparable to the degrees of agreement reported for euphemism datasets in Lee et al. (2024).

4.3 Balanced Dataset

For our text classification experiments, we sampled a portion of the main dataset. This was because some PETs had a disproportionately high number of examples compared to others, or a very skewed label imbalance (e.g., 100 euphemistic instances and 1 non-euphemistic). These factors were not ideal for text classification, and we wanted to assess models' abilities to classify texts for a variety of different PETs with different labels. Therefore, we randomly sample a maximum of 40 euphemistic and 40 non-euphemistic examples for each PET. In addition, some annotated examples, such as *apartman görevlisi* (*apartment attendant/janitor*), *inme* (*landing/paralysis*), and *toplu* (*bulk/fat*), were never used euphemistically, so we chose not to select those. The final result was a subset of 908 instances (521 euphemistic and 387 non-euphemistic) used for the euphemism detection task.

4.4 Dataset Statistics

We conducted a detailed statistical analysis of both the main and balanced datasets to better understand their differences and characteristics. Firstly, we provide the distribution of sensitive topics in Table 4. This table categorizes PETs into various groups, such as bodily functions, death, employment/finances, illness, miscellaneous, physical/mental attributes, politics, sexual activity, substances, and social topics. Each category is accompanied by the count of entries and examples of PETs within that category. Table 5 further highlights key metrics such as average sentences per example, number of tokens, and lexical density. Notably, we also compute an "PET ambiguity" score, which measures the degree of ambiguity, or class balance, for examples of a particular PET. For each PET, this was computed as follows:

$$1 - \frac{|N_{euph} - N_{noneuph}|}{N_{euph} + N_{noneuph}} \quad (1)$$

where N_{euph} and $N_{noneuph}$ is the number of euphemistic and non-euphemistic examples for that PET, respectively. Higher values indicate a higher degree of ambiguity. For example, if there were 5 euphemistic and 5 non-euphemistic examples of a particular PET, then it is maximally ambiguous (score = 1); if there were 10 euphemistic examples and 0 non-euphemistic, then the PET is not ambiguous at all (score = 0). We compute the average ambiguity score across all PETs in the main and balanced datasets for comparison. As expected, the main dataset has a significantly lower ambiguity score (0.076) compared to the balanced dataset (0.46), suggesting more consistent usage of terms in either euphemistic or non-euphemistic contexts and confirming that balanced dataset is better suited for the euphemisms detection task.

5 Methodology

5.1 Experiments

Since one of our goals were to extend the euphemism detection task to Turkish, classification experiments were conducted. Therefore, transformer-based models pre-trained on Turkish text like BERT-base-turkish-cased and ELECTRA-base-turkish-cased-discriminator were chosen due to their capability of capturing and understanding the linguistic nuances.

The balanced dataset described in the previous section was then randomly split into training (80%), testing (10%), and validation (10%) sets, resulting in 726 examples for training and 91 examples each for testing and validation. The 80-10-10 split is a common practice in machine learning for dividing a dataset into training, validation, and testing sets.

The fine-tuning process involved training each model on our prepared dataset for a maximum of 30 epochs with a learning rate of 1e-5 and a batch size of 4. We employed early stopping with a patience of 5 to prevent overfitting. No layers were frozen during fine-tuning, allowing the models to adapt fully to the euphemism detection task. Hyperparameter optimization was not explicitly performed in this initial exploration; however, the chosen hyperparameters are common for fine-tuning BERT-based models. The primary metric for evaluating model performance during training and validation was the macro-averaged F1 score, a balanced measure of precision and recall that is suitable for binary classification tasks with potentially imbalanced classes. The fine-tuned models were then evaluated on the held-out test sets, and their performance was assessed using various metrics, including accuracy, precision, recall, and F1 score.

Category	Count	PET Examples
bodily functions	2244	*sulamak* (to water), *aybaşı* (month's beginning), *hacet görmek*(to meet the need)
death	2564	*kaybetmek* (to lose), *vefat etmek* (pass away), *aramızdan ayrıldı* (left us)
employment/finances	276	*yoksul* (to be lacking), *ekonomik* (economical), *ihtiyaç sahibi* (in need)
illness	8	*amansız hastalık* (relentless disease), *ince hastalık* (thin disease)
misc.	10	*iyi saatte olsunlar* (may they be in a good hour)
physical/mental attributes	627	*görme engelli* (visually impaired), *işitme engelli* (hearing impaired)
politics	26	*sığınmacı* (seeking asylum), *gelişmekte olan ülke* (developing country)
sexual activity	190	*seks işçisi* (sex worker), *mercimeği fırına vermek* (put the lentils in the oven)
substances	143	*madde* (subtance)
social	27	*sıkmak* (to squeeze)

Table 4: Sensitive Topics with PET examples

5.2 Results

We gathered the results of all the test sets of each model and calculated the average of 20 trials (different train-validation-test splits). The findings demonstrated that monolingual models (BERT-base-turkish-cased and ELECTRA-base-turkish-cased-discriminator) outperformed the multilingual models (BERT-base-multilingual-cased and XLM-RoBERTa). This suggests that for automatic euphemism detection in Turkish, models specifically pre-trained on Turkish text data have an advantage due to their familiarity with the nuances of the language.

Additionally, the ELECTRA architecture appears to be slightly more effective for this task than the BERT architecture, as evidenced by the higher scores of ELECTRA-base-turkish-cased-discriminator compared to BERT-base-turkish-cased. This could be attributed to the discriminator's ability to better distinguish between real and fake input data during training, which might be beneficial in identifying the subtle differences between euphemistic and non-euphemistic expressions. The results obtained from the models can be seen in Table 4.

The findings of this research have several potential real-world applications. The developed models could be integrated into NLP tools for automatic euphemism detection in various types of text data, including social media posts, news articles, and other online content. This could be particularly valuable in fields such as social media monitoring to analyze the insight into public sentiment, opinions, and attitudes towards sensitive topics. For content moderation, flagging potentially harmful or offensive content that uses euphemisms to disguise its true intent could be beneficial for online platforms and communities seeking to maintain a respectful and safe environment.

Moreover, the cross-lingual capabilities of the models demonstrated in this study open up possibilities for developing euphemism detection systems for low-resource languages, where labeled data might be limited. This could contribute to a more inclusive and equitable representation of different languages and cultures in NLP research and applications.

6 Conclusion and Future Work

In this study, we created a Turkish PETs dataset from scratch and through utilizing the dataset, we investigated the effectiveness of various language models in identifying and categorizing euphemisms in Turkish. Our findings indicate that models trained on multilingual data, particularly XLM-RoBERTa, generally outperform monolingual models, suggesting the benefits of cross-lingual transfer learning in capturing euphemistic nuances. However, for the Turkish language specifically, models trained on Turkish text data, such as BERT-base-turkish-cased and ELECTRA-base-turkish-cased-

Metric	Main Dataset	Balanced Dataset
Total Examples	6115	908
Euphemistic Examples	1876	521
Non-Euphemistic Examples	4239	387
Avg. PET Ambiguity	**0.076**	**0.46**
Avg. Sentences per Example	3.60	3.28
Avg. Sentences (Euphemistic)	3.51	3.16
Avg. Sentences (Non-euphemistic)	3.63	3.43
Avg. Number of Tokens per Example	96.22	90.42
Avg. Number of Unique Tokens per Example	78.63	74.24
Avg. Lexical Density	0.82	0.84
Notable PETs (Only Non-euphemistic Examples)	18 PETs (e.g., *toplu/bulk*, *işini bitirmek/to finish his/her job*, *inme/landing*)	1 PET (*e.g. muhtaç/in need*)

Table 5: Comparison of Main and Balanced Datasets

	Accuracy	F1	Precision	Recall
mBERT	0.81	0.80	0.80	0.80
XLM-RoBERTa	0.82	0.82	0.82	0.81
BERTurk	0.84	0.84	0.84	0.84
ELECTRA	0.86	0.86	0.86	0.86

Table 6: Performance of the models on the Turkish euphemisms.

discriminator, demonstrated superior performance, emphasizing the importance of language-specific training for this task.

Future research could investigate the impact of model size, architecture, and training data on euphemism detection performance. Additionally, exploring the use of explainability techniques could provide valuable insights into the decision-making processes of these models to better comprehend the specific linguistic features they rely on for euphemism detection. Experimenting with different model architectures or training techniques might also further improve the performance of euphemism detection systems in Turkish. Additionally, expanding the dataset to include a wider range of euphemisms and exploring their application in downstream tasks like sentiment analysis and content moderation could be useful for future work. It is important to acknowledge that the results are based on a limited dataset and may not generalize to all types of euphemisms in Turkish. Future work could involve testing the models on a larger and more diverse dataset to confirm these findings.

Lastly, exploring the cross-lingual transferability of euphemism detection models trained on Turkish data to other languages, similar to the work done in Lee et al. (2023, 2024) would provide valuable insights. This could involve fine-tuning multilingual models on Turkish euphemisms and evaluating their performance on other languages. As highlighted in Gavidia et al. (2022a), the ambiguity of potentially euphemistic terms (PETs) is a major challenge; therefore, future work could focus on developing methods to disambiguate PETs and distinguish between their euphemistic and non-euphemistic usages more effectively.

Limitations

While this study highlights the potential of language models in euphemism detection in Turkish, the results are based on a limited dataset that may not encompass the full spectrum of euphemistic language usage in Turkish, potentially affecting the generalizability of our findings.

Ethics Statement

The authors foresee no ethical concerns with the work presented in this paper.

Acknowledgments

Thanks to the annotators, whose names are Kader Teke, Devran Sarısu, Sümeyye Sena Şahin, Fitnat Filiz Bal, Kübra Aksoy, Ecem Küçükler, Azra Almira Kılıç, Özge Bilik, Mihriban Kandemir, Nazan Demir, Şüheda Nur Ünal, Özlem Özer, Salih Hamza Küpeli it was possible for us to create this dataset quickly.

This material is based upon work supported by the National Science Foundation under Grant No. 2226006.

References

Facebook AI. 2019. Unsupervised cross-lingual representation learning at scale. https://huggingface.co/xlm-roberta-base.

Google AI. 2018. Multilingual bert: A universal language model. https://huggingface.co/google/bert-base-multilingual-cased.

Doğan Aksan. 1994. Göktürk yazitlarinda söz sanatlari güçlü anlatim yollari. *Türk Dili Araştırmaları Yıllığı-Belleten*, 38(1990):1–12.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Dildora Oktamovna Bakhriddionova. 2021. The needs of using euphemisms. *Mental Enlightenment Scientific-Methodological Journal*, 2021(06):55–64.

Fatih Beyhan, Buse Çarık, Inanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A turkish hate speech dataset and detection system. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 4177–4185.

Arzu ÇĠFTOĞLU Çabuk. 2015. Türkçedekġ örtmece sözlerġn oluġum yollari. *Manas Journal of Social Studies*, 4(5):136–160.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Bernard Comrie. 1988. Linguistic typology. *Annual Review of Anthropology*, 17(1):145–159.

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.

DBMDZ. 2019. Berturk: Bert models for turkish. https://huggingface.co/dbmdz/bert-base-turkish-cased.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Christian Felt and Ellen Riloff. 2020. Recognizing euphemisms and dysphemisms using sentiment analysis. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 136–145.

Martha Gavidia, Patrick Lee, Anna Feldman, and JIng Peng. 2022a. CATs are fuzzy PETs: A corpus and analysis of potentially euphemistic terms. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2658–2671, Marseille, France. European Language Resources Association.

Martha Gavidia, Patrick Lee, Anna Feldman, and Jing Peng. 2022b. CATs are Fuzzy PETs: A Corpus and Analysis of Potentially Euphemistic Terms. In *Proceedings of the 13th Language and Resources Conference*. ELRA.

Aslı Göksel and Celia Kerslake. 2004. *Turkish: A comprehensive grammar*. Routledge.

Ferhat Karabulut and Gulmira Ospanova. 2013. Örtmece sözlerin mantığı: Kazak türkçesi ile türkiye türkçesinde karşılaştırmalı model analizi. *Uluslararası Türkçe Edebiyat Kültür Eğitim (TEKE) Dergisi*, 2(2):122–146.

Savo Karam. 2011. Truths and euphemisms: How euphemisms are used in the political arena. 17.

Sedrick Scott Keh, Rohit Bharadwaj, Emmy Liu, Simone Tedeschi, Varun Gangal, and Roberto Navigli. 2022. EUREKA: EUphemism recognition enhanced through knn-based methods and augmentation. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 111–117, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ilker Kesen, Aykut Erdem, Erkut Erdem, and Iacer Calixto. 2022. Detecting euphemisms with literal descriptions and visual imagery. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 61–67, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Patrick Lee, Alain Chirino Trujillo, Diana Cuevas Plancarte, Olumide Ojo, Xinyi Liu, Iyanuoluwa Shode, Yuan Zhao, Anna Feldman, and Jing Peng. 2024. MEDs for PETs: Multilingual euphemism disambiguation for potentially euphemistic terms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 875–881, St. Julian's, Malta. Association for Computational Linguistics.

Patrick Lee and Anna Feldman. 2024. Multilingual euphemism detection shared task: Fourth workshop on figurative language processing. https://msuweb.montclair.edu/~feldmana/.

Patrick Lee, Anna Feldman, and Jing Peng. 2022a. A report on the euphemisms detection shared task. *arXiv preprint arXiv:2211.13327*.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022b. Searching for PETs: Using distributional and sentiment-based methods to find potentially euphemistic terms. In *Proceedings of the Second Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Seattle, USA. Association for Computational Linguistics.

Patrick Lee, Martha Gavidia, Anna Feldman, and Jing Peng. 2022c. Searching for pets: Using distributional and sentiment-based methods to find potentially euphemistic terms. *Preprint*, arXiv:2205.10451.

Patrick Lee, Iyanuoluwa Shode, Alain Trujillo, Yuan Zhao, Olumide Ojo, Diana Plancarte, Anna Feldman, and Jing Peng. 2023. FEED PETs: Further experimentation and expansion on the disambiguation of potentially euphemistic terms. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 437–448, Toronto, Canada. Association for Computational Linguistics.

Abulimiti Maimaitituoheti, Yang Yong, and Fan Xiaochao. 2022. A prompt based approach for euphemism detection. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 8–12, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mustafa Melih Mutlu and Arzucan Özgür. 2022. A dataset and BERT-based models for targeted sentiment analysis on Turkish texts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 467–472, Dublin, Ireland. Association for Computational Linguistics.

Steven Pinker. 1994. The Game of the Name. *The New York Times*.

Steven Pinker. 2003. *The Blank Slate: The Modern Denial of Human Nature*. Penguin.

Hussein Rababah. 2014. The translatability and use of x-phemism expressions (x-phemization): Euphemisms, dysphemisms and orthophemisms) in the medical discourse. *Studies in Literature and Language*, 9:1–12.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Taner Sezer. 2017. Ts corpus project: An online turkish dictionary and ts diy corpus. *European Journal of Language and Literature*, 9:18.

Wanzheng Zhu and Suma Bhat. 2021. Euphemistic phrase detection by masked language model. *Preprint*, arXiv:2109.04666.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. *arXiv preprint arXiv:2103.16808*.