Detecting Hate Speech in Turkish Print Media: A Corpus and A Hybrid Approach with Target-oriented Linguistic Knowledge

Gökçe Uludoğan[1] and **Atıf Emre Yüksel**[1] and **Ümit Can Tunçer**[2]
Burak Işık [2] and **Yasemin Korkmaz**[3] and **Didar Akar**[2] and **Arzucan Özgür**[1]
[1]Department of Computer Engineering, Bogazici University, Istanbul, Turkey 34342
[2]Department of Linguistics, Bogazici University, Istanbul, Turkey 34342
[3] Hrant Dink Foundatio, Istanbul, Turkey 34373
{gokce.uludogan,akar,arzucan.ozgur}@bogazici.edu.tr

Abstract

The use of hate speech targeting ethnicity, nationalities, religious identities, and specific groups has been on the rise in the news media. However, most existing automatic hate speech detection models focus on identifying hate speech, often neglecting the target group-specific language that is common in news articles. To address this problem, we first compile a hate speech dataset, TurkishHatePrintCorpus, derived from Turkish news articles and annotate it specifically for the language related to the targeted group. We then introduce the HateTargetBERT model, which integrates the target-centric linguistic features extracted in this study into the BERT model, and demonstrate its effectiveness in detecting hate speech while allowing the model's classification decision to be explained. We have made the dataset and source code publicly available at https://github.com/boun-tabi/HateTargetBERT-TR.

Warning: This paper contains hate speech and offensive terms directed towards specific groups.

1 Introduction

Hate speech, typically characterized by defamatory statements targeted at specific groups based on ethnicity, nationality, religion, color, gender, sexual orientation, among other characteristics (Schmidt and Wiegand, 2019), presents unique challenges in media discourse. Contrary to the expectation of objectivity in news and print media, hate speech is surprisingly prevalent(HDF Publications, 2019). This study explores this phenomenon, broadening the scope to include discriminatory speech, which, while not explicitly hateful, still fosters discrimination. Despite regulatory efforts, such speech persists in media, often masked by subtle linguistic tactics. For example, distortion involves making unfair generalizations, as seen in headlines like "Greeks deliberately target refugees on sinking boat." Similarly, symbolization uses identity traits to convey messages, evident in phrases like "Will a Muslim represent us at Eurovision?" These methods not only spread hate speech but also magnify its damaging effects, highlighting the need for vigilant monitoring and action.

With the significant advances in pre-trained large language models and the transformer architecture in natural language processing, researchers have developed various architectures based on BERT (Devlin et al., 2019) that have achieved successful results in the area of hate speech detection (Mozafari et al., 2019; Gupta et al., 2020; Mozafari et al., 2020; Caselli et al., 2021; Perifanos and Goutsos, 2021). Although lexical and linguistic features have been used in different model architectures (Nobata et al., 2016; Wiegand et al., 2018; Koufakou et al., 2020; Hüsünbeyi et al., 2022), the integration of target-oriented linguistic features into the BERT model has not yet been studied.

There are open data sets on certain aspects of hate speech in different languages and especially in social media (Zampieri et al., 2019; Basile et al., 2019; Sap et al., 2020; ElSherief et al., 2021). However, resources for languages like Turkish are scarce (Mayda et al., 2021). Recent studies have addressed this issue by compiling Turkish tweets from "hate domains" on specific topics like politics, religion, and vaccination where hate speech might emerge (Beyhan et al., 2022; Arın et al., 2023; İhtiyar et al., 2023). Concurrently, BERT-based models are being developed for hate speech detection (Toraman et al., 2022; Beyhan et al., 2022). Previous work has also focused on hate speech in Turkish news articles and proposed a hybrid model for hate speech detection by integrating linguistic features into BERT (Hüsünbeyi et al., 2022). However, the linguistic features used in this study rely on general morpho-syntactic properties of Turkish, neglecting the crucial aspect of the target groups of hate speech. In this study, we compile a dataset of hate speech derived from Turkish print news and

Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 205–214
March 22, 2024 ©2024 Association for Computational Linguistics

annotate it specifically for language related to the targeted group. We then introduce the HateTargetBERT model, which integrates the target-centric linguistic features extracted in this study into the BERT model, and demonstrate its effectiveness in detecting hate speech while allowing the model's classification decision to be explained.

The main contributions of this paper can be summarized as follows: (i) We develop HateTargetBERT, a model that couples BERT with hate speech target-oriented linguistic features extracted from hate speech content in the news articles and enables the generation of an explanation for the model's classification decision. (ii) We release `TurkishHatePrintCorpus`, a human-annotated hate speech dataset derived from Turkish print media and make the dataset, our model, and its source code publicly available[1].

2 Dataset

2.1 Collection

To compile a dataset of newspaper articles containing hate speech, we collected articles from various Turkish print media outlets. These articles were selected based on specific keywords associated with the target groups such as ethnicity, nationality, and religious identity. The keywords we used for querying were selected by the linguists in our team based on a combination of domain knowledge and an initial exploration of the print media. We aim to capture a wide range of hate speech instances in the Turkish print media context. The printed articles, initially in the form of scanned images, were obtained from PRNet, a company that provides a media archive and an OCR tool. We used this OCR tool to convert the scanned images into text format.

2.2 Filtering

Collecting articles from print media presents unique challenges. Many of these articles contain Optical Character Recognition (OCR) errors at both word and sentence levels. Instances have been observed where sentences are distorted as a result of the joining of two half-sentences from double-column printing. To enhance data quality, we adopted a filtering strategy that relies on scoring words and sentences using an n-gram language model. To achieve this, articles were segmented into sentences and tokenized using the Zemberek library[2]. Both the sentences and words were then scored employing a 5-gram model, which was trained with the KenLM library[3] on a recent dump of the Turkish Wikipedia using subword tokenization. The scoring process incorporated length-based normalization to facilitate fair comparisons. We calculated the mean and standard deviation of sentence scores for each article. A manual analysis was performed on both sentences and words to establish thresholds for anomalies. Next, we computed the ratio of anomalous words and sentences within an article. To refine the collected articles, we applied the following criteria:

- Articles shouldn't contain sentences that score less than -1.9 using a language model, indicating they are anomalous.
- The average proportion of anomalous tokens in a sentence should not exceed 20%.
- No sentence within the article should have an anomalous token ratio greater than 50%.
- On average, a sentence in an article should have 2 or fewer anomalous tokens.
- The mean score for the sentences in the article should be greater than -0.61.
- Sentence scores within an article should have a standard deviation below 0.2.

Additionally, we filtered content at the article level. During preprocessing, we removed URLs, emails, numbers, currency symbols, and non-Turkish words using the langdetect library[4].

2.3 Annotation

The annotation process involved both volunteers and a project team. These volunteers were predominantly university students from diverse fields, including media studies and sociology. Their selection was based on both their expressed interest in the topic and a review of their resumes. Before the annotation, we ensured that the volunteers underwent a comprehensive training session. In this session, they were introduced to our definition of hate speech: statements that marginalize, threaten, or insult groups based on their ethnicity, nationality, or religious identity. Notably, this definition

[1] https://github.com/boun-tabi/HateTargetBERT-TR

[2] https://github.com/loodos/zemberek-python

[3] https://github.com/kpu/kenlm

[4] https://github.com/Mimino666/langdetect

excludes comments directed at individual persons, institutions, or organizations.The analysis of articles that mention ethnic, national, or religious groups is guided by key questions: Following the clarification of various hate speech categories, the texts containing hate speech are discussed in relation to categories. To enhance their understanding, they were provided with representative examples from the print media. Several examples of hate speech expression in the news articles can be found in Table 1.

Each volunteer worked independently, identifying articles containing hate speech and marking those that were ambiguous. Once a day's articles were annotated, they were collectively reviewed with the project team. During this review process, any contradictory content within the articles sparked methodological and conceptual debates. Through collaborative discussions, the volunteers and project team achieved consensus on the article annotations. To validate the annotations, secondary annotators reviewed ten percent of the randomly selected articles, resulting in a Cohen's Kappa score of 0.675, indicating substantial agreement between annotators. Upon identifying newspaper articles containing hate speech, we selected one non-hateful newspaper article from the same day for each hateful newspaper article.

2.4 Statistics

Compiled from 859 distinct media sources, `TurkishHatePrintCorpus` provides an extensive scope for analyzing the linguistic characteristics and distinctions between hateful and nonhateful articles. The dataset displays the variety in the number of articles collected from each source. While we obtained only one article from 274 outlets, a significant portion of the corpus is supported by the prominent contributions of a few outlets. Notably, the top five outlets from which we gathered articles contributed 299, 205, 159, 155, and 143 articles, respectively.

The dataset comprises 3406 articles from local media sources along with 3275 articles from national ones. As for the hate speech categories, `TurkishHatePrintCorpus` contains 3678 nonhateful articles along with 3003 hateful ones.

Each article in the dataset, on average, comprises around 21 sentences. Articles in the dataset vary, with some being as brief as 2 sentences and others as lengthy as 263 sentences. Moreover, the average word count for an article stands at 350 words, with some articles having as few as 21 words and others boasting a word count as high as 3047.

Table 2 presents an overview of the general statistics for this annotated dataset, while Table 3 details the distribution of news articles with hate speech and the corresponding target groups.

The curated dataset was then divided into training, validation and test sets, ensuring that the ratio of hateful to non-hateful news articles remained consistent across all sets. The distribution of hateful and non-hateful newspaper articles across the splits is shown in Table 4.

3 Methodology

We develop HateTargetBERT, a model that couples BERT with target-oriented linguistic features specifically designed for hate speech detection. As illustrated in Figure 1, the model architecture consists of a BERT model followed by fully connected network (FCN) layers. These layers not only take the last hidden representation of the `[CLS]` token, which is typically used as a sentence embedding, but also incorporate the extracted linguistic features as input. To prevent overfitting, we incorporate dropout layers (Srivastava et al., 2014) between the FCN layers.

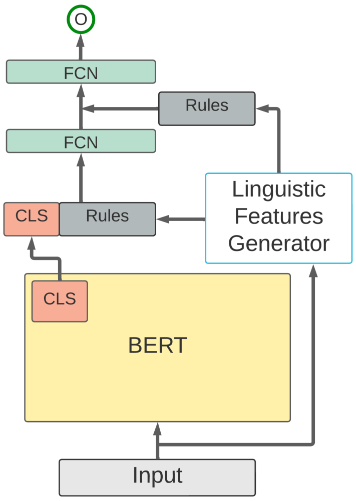

Figure 1: Overview of HateTargetBERT.

3.1 Linguistic Features

In HateTargetBERT, linguistic features serve as additional indicators for hate speech detection. These features focus on target groups, hateful words, ethnicity-specific rules, and unique pat-

Table 1: Examples of hate speech expression in the news articles

	Target	Type
...Bin mülteciye bakamayan Yunanlılar onları şiddet ve dayakla Türkiye'ye göndermeye devam ediyor... (...Greeks who cannot take care of a thousand refugees continue to send them to Turkey with violence and beatings...)	greek	hostility/war discourse
...Avrupalılar önce terörü üretti. Sonra güya kendileri mücadele ediyor... (...Europeans first produced terror. Then they are supposedly struggling themselves...)	european	exaggeration/attribution/distortion
...Böylesine zulmü gavur bile yapmadı... (...Even infidels did not commit such cruelty...)	infidel	symbolization

Table 2: Statistics of the human-annotated hate-speech print media dataset.

Statistics	Detail
Number of samples	6681
Number of sources	859
Articles from local sources / national sources	3406 / 3275
Time period	2014-2019
Average number of sentences per article	21
Average number of words per article	350

terns that identify hate speech for a specific target group. Linguists in our team derived these features by utilizing the trTenTen corpus [5] available on SketchEngine, using the names of target groups as keywords, to find patterns potentially indicative of hate speech.

The linguistic features are grouped into five categories (i.e., types), each with unique characteristics in terms of feature formulation, hate speech content search methodology, and semantic expression. A summary of these features is presented in Table 5. Each category, except the target agnostic type, is further divided into several subtypes based on the severity of hate speech, as determined by linguistic experts. The severity ranges from Degree 1 (least severe) to Degree 5 (most severe). Each feature is represented with one-hot encoding, except for those of target agnostic type, which accumulate the number of detected rules. Some feature types are searched in a range of window while others require strict matches.

Target-agnostic features aim to identify patterns common across all ethnicities and nationalities in news articles, using a variety of terms often found in hate speech. These patterns are searched within a 15-word range (see Table 6 for a list of patterns). These patterns were developed considering Turkish grammar, an agglutinative language with a Subject-Object-Verb structure where nouns take suffixes based on their role. For example, if a noun

Table 3: Number of occurrences of hate speech target groups in hateful and non-hateful articles within `TurkishHatePrintCorpus`.

Target	Hateful / Non-hateful	Target	Hateful / Non-hateful
Afghan	119 / 104	Immigrant	169 / 214
Alevi	25 / 82	Infidels	65 / 4
Arab	336 / 223	Iranian	20 / 24
Armenian	847 / 140	Iraqi	27 / 29
Assyrian	14 / 13	Italian	75 / 45
Atheist	36 / 4	Jewish	25 / 20
Buddhist	112 / 10	Kurdish	265 / 181
Bulgarian	61 / 46	Kyrgyz	12 / 11
Catholic	35 / 11	Lebanese	7 / 5
Chechen	12 / 4	Muslim	886 / 593
Chinese	17 / 17	Orthodox	32 / 11
Christian	372 / 128	Pakistani	65 / 68
Crusader	149 / 56	Refugee	265 / 374
Dutch	29 / 15	Russian	665 / 468
English	336 / 142	Saudi	118 / 80
European	117 / 65	Serbian	83 / 15
French	230 / 98	Syrian	646 / 555
German	348 / 307	Turbaned	3 / 1
Giaour	32 / 6	Turkmen	60 / 63
Greek (Rum)	799 / 728	Ukranian	1 / 8
Greek (Yunan)	541 / 379	Western	255 / 174
Gypsies	8 / 9	Yazidi	27 / 18
Hebrew	646 / 142	Yemeni	8 / 4
Hungarian	31 / 38		

[5] https://www.sketchengine.eu/trtenten-turkish-corpus

Table 4: Number of samples in each class across data splits.

Split	Hateful	Non-hateful
Training	2395	2949
Validation	305	363
Test	303	366

from a target group is near the active verb "öldür-" (to kill) and is in the nominative form (suffix-free in Turkish), it's likely the sentence's agent. Similarly, if "tarafindan" (by) is near the passive verb "öldürül-" (to be killed), the preceding word is the agent. If this word is from the target group, it suggests that the target group is the agent. Patterns were created using fixed words and variables. Functional words like "tarafindan" (by) and suffixes such as -A (dative), -(n)In (genitive) are fixed, while target group names, adjectives, verbs, and gerunds are variable. **Target-specific features**, on the other hand, aim to detect patterns that are generally associated with a particular group in news using the same approach (see Table 7 for details).

Pre-target and post-target features are designed to identify hateful patterns that are adjacent to particular targets. These features highlight the specific hate speech content that authors aim to promote in the news. The adjacent features are identified through direct pattern matching, without the use of a window parameter. These features are categorized based on their severity, as determined by linguistic experts. For instance, a pre-target pattern like "covert `[ETHN]`" is considered to be of Degree 1 severity, indicating a less severe form of hate speech. In this pattern, `[ETHN]` serves as a placeholder representing any ethnicity. On the other hand, a post-target pattern such as "`[ETHN]` treachery" is of Degree 5 severity, indicating a more severe form of hate speech. For a comprehensive list of pre-target and post-target features , please refer to Table 8 and 9, respectively.

Misleading nonhateful patterns are patterns that appear in newspaper articles about target groups but don't typically indicate hate speech. They are identified by detecting specific word sequences around the target keyword that are likely to come from a non-hateful context. For instance, "`[ETHN]` footballer" probably originates from a sports article. Moreover, some phrases with the target group are not considered hate speech. For example, "Kürt terör örgütü" (Kurdish terrorist organization) is seen as hate speech due to its ethnic emphasis, but "Kürtçü terör örgütü" (Kurdist terrorist organization) isn't, as it emphasizes the organization's ideology. Additionally, quotes from individuals, indicated with phrases like "dedi" (he/she said), are not evaluated for hate speech in this study. A comprehensive list of these patterns can be found in Table 10.

3.2 Baseline models

We compare our model with two other models: BERTurk (Schweter, 2020), which solely leverages BERT representations, and HateTargetNN, a basic two-layer fully-connected network that only uses the linguistic features extracted in this study.
BERTurk (Schweter, 2020) is a transformer based model pretrained on a compilation of Turkish OSCAR[6], Wikipedia dump, and various OPUS corpora[7]. It has been shown to be one of the state-of-the-art models for hate speech detection in Turkish text (Hüsünbeyi et al., 2022; Beyhan et al., 2022). To adapt it for hate speech detection, we fine-tuned the pretrained model on the curated hate speech dataset by adding a fully-connected layer that utilizes the `[CLS]` token representation.
HateTargetNN is another baseline model that we use to test the ability of the linguistic features alone in detecting hate speech. This model is a two-layer fully-connected neural network, which includes batch normalization and dropout layers.

3.3 Implementation Details

We adopt BERTurk (Schweter, 2020) as the initial checkpoint for our HateTargetBERT model. All hyperparameters are selected based on their performance on the validation set. We use the F1 score as the metric to evaluate the performance of the models on the validation set, with the validation performance assessed each epoch.

We trained BERTurk and HateTargetBERT for 3 epochs while HateTargetNN is trained for 10 epochs. For the BERT-based models, we used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e-5 and a weight decay of 1e-2. Conversely, for HateTargetNN, we retained the same settings but adjusted the learning rate to 1e-3. We also incorporated a scheduler for the learning rate, with a patience of 2 evaluation steps and a

[6] https://traces1.inria.fr/oscar/

[7] http://opus.nlpl.eu/

Table 5: Summary of the target-oriented linguistic features. These features are divided into six categories: target agnostic, target specific, pre-target, post-target, and misleading non-hateful features. Each type, except for the target agnostic type, is further divided into several subtypes based on the severity of hate speech, from Degree 1 (least severe) to Degree 5 (most severe). [ETHN] substitutes for any ethnicity.

Type	# Subtypes	# Patterns	Window	Example (in Turkish)	Translation
Target agnostic	-	14		[IRK]ın kuklası olan [IRK]	[ETHN] who are puppets of [OTHER ETHN]
Target specific	4	19		kripto ermeni	crypto armenian
Pre-Target	5	38		istilacı [IRK]	invading [ETHN]
Post-Target	5	60		[IRK] soykırımı	[ETHN] massacres
Misleading Nonhateful	4	30		[IRK] filozof	[ETHN] philosopher

Table 6: Target agnostic patterns that are frequently used in hate speech content. [ETHN] substitutes for any ethnicity to generate feature. Patterns are searched in determined window range. [ADJBEF] and [ADJAFTER] indicate that words from Tables 8 and 9 can be placed respectively.

ID	Pattern	EN Translation
1	[IRK]+lık yapmak	act like [ETHN]
2	[IRK]a bak sen	look at that [ETHN]
3	[IRK [IRK]+lığını yap(mak)	s/he does her/his [ETHN]
4	[IRK] kurşunlarıyla/bombalarıyla/parasıyla	with the bullets/bombs/money of [ETHN]
5	[IRK] paryası/skandalı/işgali/baskını	[ETHN] pariah/scandal/occupation/invasion
6	[IRK]+ın gerçekleştirdiği/yaptığı katliam/zulüm/soykırım	massacre/persecution/cruelty/oppressure of [ETHN]
7	[IRK]+ın uşağı/işbirlikçisi/piyonu/kuklası (olan) [IRK]	[ETHN] servant/pawn/collaborator/puppets of [OTHER ETHN]
8	[IRK] destekli [IRK] darbesi/saldırıları/katliamı/soykırımı	[ETHN] backed [OTHER ETHN] coup/genocide/massacre/attacks
9	[IRK] tarafından saldırıya/katliama/soykırıma maruz kalmak/uğramak	being attacked/subjected to genocide/massacred by the [ETHN]
10	[IRK] tarafından gerçekleştirilen/yapılan katliam/zulüm/soykırım	massacre/persecution/cruelty/oppressure done/carried out by [ETHN]
11	[IRK] ... öldürdü/katletti/etnik temizlik yaptı/kirletti/bastı/şehit etti	[ETHN] ... killed/massacred/did ethnic cleansing/disgloried/martyrized
12	[IRK] tarafından ... öldürüldü/katledildi/etnik temizlik yapıldı/basıldı/şehit edildi	killed/massacred/did ethnic cleansing/disgloried/martyrized ... by [ETHN]
13	[IRK] tarafından IRK+a yönelik saldırılar/katliam/zulüm/soykırım	genocide/massacre/persecution/cruelty/oppressure/attack of [ETHN] by [OTHER ETHN]
14	[IRK]+ın hain(ce)/vahşi(ce)/insanlık dışı/hunharca/kan donduran/şeytani/sinsi/[ADJBEF] teşebbüsleri/planları/oluşumları/[ADJAFTER]	sneaky/traitorous/wild/subhuman/bloodthirstily/terrifical/satanic/[ADJBEF] [ETHN]'s attempts/plans/organizations/[ADJAFTER]

Table 7: Target specific patterns. Higher degree points more serious hate speech content.

Degree 1	Degree 2	Degree 3	Degree 4
vahşi toplumlar (wild societies)	batıl batı (superstitious west)	yahudi ajanı (jewish agent)	yahudi çakallığı (jewish cowardice)
batı cehaleti (western ignorance)		yahudi uşağı (jewish servant)	katil rum (killer rum)
haçlı zihniyeti (crusader mentality)		kripto ermeni (crypto armenian)	haydut rumlar (rogue Greeks)
kriptolar (cryptos)		suriyeli işgali (syrian invasion)	rum zorbalığı (Greek bullying)
		afgan işgali (afghan invasion)	kafir alevi (infidel alevist)
		mülteci işgali (refugee invasion)	ateist alevi (atheist alevist)
		pakistanlı işgali (pakistani invasion)	
		arapların işgali (invasion of the arabs)	

Table 8: Pre-target features in hate speech content. A higher degree indicates a more serious hate speech content.

Degree 1	Degree 2	Degree 3	Degree 4	Degree 5
sapıtan (amok)	katleden (murderous)	kripto (crypto)	işgalci (invader)	hain (traitorous)
çakma (fake)	korkak (coward)	sinsi (sly)	gaspçı (grabber)	katleden (murderous)
facir (sinner)	yamyam (cannibal)	açgözlü (greedy)	lanetlenmiş (damned)	gavur (infidel)
gizli (covert)	başbelası (the very devil)	dönek (renegade)	Allah'ın lanetlediği (cursed by god)	kalleş (treacherous)
kışkırmış (spoiled)		iki yüzlü (two-faced)	zalim (cruel)	kan gölüne çeviren (vicious killer)
hırsız (thief)		azgın (ferocious)	şerefsiz (dishonourable)	insanlık suçu işleyen (perpetrator of crimes against humanity)
		edepsiz (shameless)	gaddar (grim)	bebek katili (baby murderer)
		yağmacı (predatory)	gasıp (usurper)	cani (villain)
		çapulcu (marauder)	canavarlaşmış (monstrous)	vahşi (wild)
				eli kanlı (bloody) hand

reduction factor of 0.5. The dropout probability of the additional layers in HateTargetBERT was set to 0.5. It is worth noting that the models underwent training on ten unique splits, each initialized with different seeds, and were subsequently evaluated on the test set.

4 Results

As shown in Table 11, HateTargetBERT, combining BERT with target-oriented features, demonstrated superior performance compared to the baseline HateTargetNN model, which solely relies on linguistic features. Additionally, HateTargetBERT performed at a comparable level to BERTurk. Al-

Table 9: Post-target features in hate speech content. A higher degree indicates a more serious hate speech content.

Degree 1	**Degree 2**	**Degree 3**	**Degree 4**	**Degree 5**
işbirlikçisi (collaborator)	yalanları (lies)	baskısı (pressure)	terörü (terror)	gaddarlığı (atrocity)
inadı (stubbornness)	iftiraları (slanders)	bozma (violation)	terör üssü (terror base)	imha (destruction)
doyumsuzluğu (dissatisfaction)	tehdidi (threat)	yalakalığı (fawning)	saldırıları (attacks)	zulmü (cruelty)
karısı (wife)	oyunu (games)	entrikaları (intrigues)	terörizmi (terrorism)	kırımı (politicide)
dolandırıcı (swindler)	yağmacılar (looters)	fesatları (mischief)	sapkınlığı (heresy)	vahşeti (brutality)
parmağı (hand)	çapulcusu (marauder)	sürüleri (herds)	köpekler (dogs)	zalimi (ferocity)
provokasyonu (provocation)	haydutlar (bandits)	kötülükleri (evil)	terör örgütü (terrorist organization)	hain (traitor)
artığı (reversion)		dönekliği (apostasy)	sırtlanlar (hyenas)	kalleşliği (treachery)
uşaklığı (servitude)		açgözlülüğü (greed)	soysuzlar (retrograde)	canilikleri (murderousness)
aşığı (lover)		sinsiliği (snakiness)	çakallar (coyotes)	kıyımları (massacres)
gaspçılar (usurpers)		yüzsüzlüğü (sassiness)	yamyamlar (cannibals)	piçleri (bastards)
kuklası (puppets)		iki yüzlülüğü (hypocrisy)	vandallar (vandals)	
piyonları (pawns)		baskını (raid)		
teröristi (terrorist)		tohumu (seed)		
virüsü (virus)		dölleri (spawn)		
sevici (lover)				

Table 10: Misleading hate speech content that are found mostly non hate speech news. [ETHN] substitutes for any ethnicity to generate a feature. A higher rating indicates less or no hate speech. "[ETHN]+ist" expresses ethnicity names and the suffix"-CU" in Turkish, which is derived nationalist names from it by attached them (e.g. Türkçü, Kürtçü).

Degree 1	**Degree 2**	**Degree 3**	**Degree 4**
[IRK] çeteleri ([ETHN] gangs)	haçlı seferi (crusade)	diye belirtti (s/he stated)	futbol (football)
[IRK] fanatiği ([ETHN] fanatics)	STK (non-governmental organizations)	dedi (said)	spor (sport)
[IRK]+cı terör örgütü ([ETHN]+ist terrorist organization)	tarihte bugün (today in history)	şeklinde açıkladı (expressed as)	maç (match)
[IRK] polisi ([ETHN] police)	takvimde bugün (today on the calendar)	şeklinde ifade etti (explained as)	antik yunan (ancient greek)
[IRK] yaygaracılığı ([ETHN] fuss)			yunan düşünür (greek thinker)
[IRK] askerleri ([ETHN] soldiers)			yunan filozof (greek philosopher)
[IRK] milisleri ([ETHN] militia)			
[IRK] militanları ([ETHN] militants)			
[IRK] yerleşimciler ([ETHN] settlers)			
[IRK] milliyetçiler ([ETHN] nationalists)			
[IRK] güçleri ([ETHN] forces)			
[IRK] isyanı ([ETHN] revolt)			
radikal [IRK] (radical [ETHN])			
ırkçı [IRK] (racist [ETHN])			
siyonistler (zionist jew)			
pontus rum (pontus empire)			

Table 11: Evaluation of the models on the test set .

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
HateTargetNN	66.38 ±0.89	83.66 ±4.45	31.39 ±2.24	45.62 ±2.75
BERTurk (Schweter, 2020)	90.60 ±1.20	87.69 ±2.49	92.02 ±2.15	89.78 ±1.53
HateTargetBERT	90.54 ±0.84	88.47 ±2.18	90.82 ±2.07	89.60 ±1.16

Table 12: Evaluation of the models on the test instances with at least one linguistic feature.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
HateTargetNN	75.22 ±2.10	83.70 ±3.88	58.57 ±4.47	68.79 ±3.45
BERTurk (Schweter, 2020)	90.49 ±1.92	88.39 ±3.59	91.80 ±1.90	90.03 ±2.15
HateTargetBERT	90.75 ±1.08	89.37 ±2.75	91.19 ±2.29	90.22 ±1.24

though BERTurk exhibited slightly better scores across various metrics, except for precision, the differences were not statistically significant based on the two-tailed paired t-test conducted at a 95% confidence interval. It is important to note that while the linguistic features were applied to all instances, only a subset of test instances contained these features, limiting their coverage. Table 12 presents a comparison of model performances on these specific instances. Notably, the models utilizing target-oriented features achieved higher scores across metrics in this subset, suggesting the effectiveness of these features and emphasizing the need for a comprehensive feature set.

We also conducted a user study using the Qualtrics online survey tool[8] to demonstrate the effectiveness of the HateTargetBERT model. For this purpose, we randomly selected ten articles from the test set that were predicted to contain hateful content and asked participants to rank the linguistic features shown in the articles based on their helpfulness in understanding the model's prediction of hatefulness. Each article was rated on a 5-point Likert scale (Strongly agree = 5, Somewhat agree = 4, Neither agree nor disagree = 3, Somewhat disagree = 2, Strongly disagree = 1). Table 13 illustrates an excerpt from a sample article from the user study.

Table 13: Excerpt from the user study and its English translation where "rum sevici" (Greek-loving) is highlighted as a post-target feature of degree 1.

Article Excerpt
... bazı önemli milliyetçi şahsiyetler kişisel menfaatlerine hizmet edilmediği değerlendirmeleriyle gidip bir kez daha Akıncı veya benzeri teslimiyetçi ve rum sevici bir başka adaya, sırf inat olsun diye oy vererek göreve getirecekler ...
English Translation
... some significant nationalist figures, assessing that their personal interests are not served, will once again go and, just out of spite, vote for another candidate like Akıncı or a similar defeatist and Greek-loving , bringing them into office ...

The study involved 25 participants, all of whom hold at least a higher education degree. The responses, as shown in Figure 2, had an average score of 3.41 and a standard deviation of 1.24. The majority of these responses fell into the categories of "strongly agree" or "somewhat agree", suggesting that the linguistic features were helpful in understanding the model's choice.

5 Related Work

Automated detection of hate speech has been extensively studied over the years due to its positive impact on society. Many studies have proposed methods to identify hateful content across different platforms in order to assist in content moderation. Previous work utilized traditional machine learning models and neural networks that leveraged features such as tf-idf, word vectors (Saha et al., 2018; de Andrade and Gonçalves, 2021), n-grams (Nobata et al., 2016; Waseem and Hovy, 2016), and lexical features (Warner and Hirschberg, 2012; Wiegand et al., 2018; Capozzi et al., 2019; Koufakou et al., 2020). However, more recent models have developed various architectures based on BERT, resulting in significant performance improvements (Mozafari et al., 2019; Gupta et al., 2020; Mozafari et al., 2020; Caselli et al., 2021; Perifanos and Goutsos, 2021).

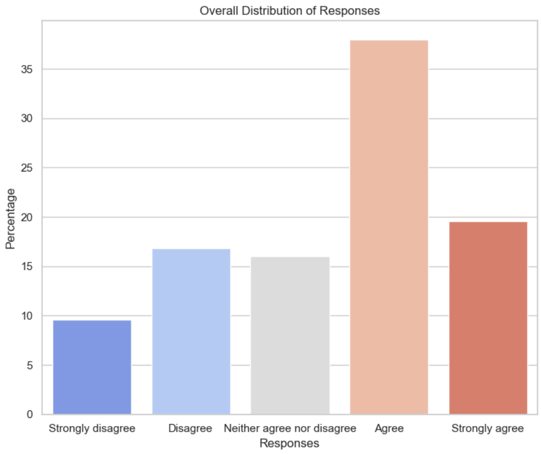

Figure 2: Distribution of participant's responses

Although hate speech is a topic that has attracted a lot of attention, there is a lack of resources for languages like Turkish (Mayda et al., 2021). Recently, several datasets have been compiled from Turkish tweets (Beyhan et al., 2022; Arın et al., 2023; İhtiyar et al., 2023). Concurrently, BERT-based models are being developed to detect hateful content in these tweets (Toraman et al., 2022; Beyhan et al., 2022). Previous work has focused on hate speech in Turkish news articles and proposed a hybrid model for hate speech detection by integrating linguistic features into BERT (Hüsünbeyi et al., 2022). However, the linguistic features used in this study rely on general morpho-syntactic properties of Turkish, neglecting the crucial aspect of the target groups of hate speech. In our work, we address this challenge by building a model that combines target-centric linguistic features with BERT. This approach achieves high performance while also providing explainability, which is particularly im-

[8] https://www.qualtrics.com

portant when dealing with longer contexts such as news articles.

6 Conclusion

We introduced TurkishHatePrintCorpus, a manually annotated hate speech dataset, compiled from Turkish newspaper articles and categorized for target groups. In addition, we developed a model, HateTargetBERT, combining BERT with target-oriented linguistic features. The results demonstrate that integrating target-oriented linguistic knowledge into a transformer model is an effective strategy for hate speech detection and for the explanation of the model's classification decision.

Limitations

This study focuses on print media, excluding the less formal and more explicit language often found in social media. Therefore, the targeted linguistic feature set are derived from printed newspaper articles. Additionally, this work aims to detect hate speech against ethnicity, national and religious entities, and immigrants. As such, newspaper articles associated with other hate domains, such as gender, are not considered. It's also worth noting that some patterns in the targeted linguistic features might be unique to Turkish. Another limitation of this study is the model's inability to handle long context lengths, exceeding 512 tokens, a common occurrence in column articles.

Ethical Considerations

We acknowledge the potential risk associated with releasing our source code and the manually annotated hate speech dataset. However, we believe that the benefits of automatic hate speech detection outweigh the associated risks of releasing the code and the dataset.

Acknowledgments

This research was partially supported by the Swedish Consulate-General, İstanbul Turkey (Project number: UM2021/10687/ISTA).

References

İnanç Arın, Zeynep Işık, Seçilay Kutal, Somaiyeh Dehghan, Arzucan Özgür, and Berrin Yanikoğlu. 2023. SIU2023-NST- Hate speech detection contest. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Fatih Beyhan, Buse Çarık, İnanç 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, Marseille, France. European Language Resources Association.

Arthur TE Capozzi, Mirko Lai, Valerio Basile, Fabio Poletto, Manuela Sanguinetti, Cristina Bosco, Viviana Patti, Giancarlo Ruffo, Cataldo Musto, Marco Polignano, et al. 2019. Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project. In *CEUR Workshop Proceedings*, volume 2481, pages 1–6. CEUR-WS.

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.

Claudio Moisés Valiense de Andrade and Marcos André Gonçalves. 2021. Profiling hate speech spreaders on twitter: Exploiting textual analysis of tweets and combinations of multiple textual representations. In *CEUR Workshop Proc*, volume 2936, pages 2186–2192.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Shailja Gupta, Sachin Lakra, and Manpreet Kaur. 2020. Study on BERT model for hate speech detection. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1–8.

Hrant Dink Foundation HDF Publications. 2019. Hate speech and discriminatory discourse in media 2019.

Zehra Melce Hüsünbeyi, Didar Akar, and Arzucan Özgür. 2022. Identifying hate speech using neural networks and discourse analysis techniques. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 32–41, Marseille, France. European Language Resources Association.

Musa İhtiyar, Ömer Özdemir, Mustafa Erengül, and Arzucan Özgür. 2023. A dataset for investigating the impact of context for offensive language detection in tweets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1543–1549.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

İslam Mayda, Yunus Emre Demir, Tuğba Dalyan, and Banu Diri. 2021. Hate speech dataset from Turkish tweets. In *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, 15(8):e0237861.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction*, 5(7):34.

Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:1812.06700*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2019. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain*, pages 1–10. Association for Computational Linguistics.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *NAACL*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

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.