Multi-domain Hate Speech Detection Using Dual Contrastive Learning and Paralinguistic Features

Somaiyeh Dehghan[1,2], **Berrin Yanikoglu**[1,2]
[1]Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey 34956
[2]Center of Excellence in Data Analytics (VERIM), Sabanci University, Istanbul, Turkey 34956
{somaiyeh.dehghan, berrin}@sabanciuniv.edu

Abstract

Social networks have become venues where people can share and spread hate speech, especially when the platforms allow users to remain anonymous. Hate speech can have significant social and cultural effects, especially when it targets specific groups of people in terms of religion, race, ethnicity, culture or a specific social situation such as immigrants and refugees. In this study, we propose a hate speech detection model, BERTurk-DualCL, using a mixed objective with contrastive learning loss that is combined with the traditional cross-entropy loss used for classification. In addition, we study the effects of paralinguistic features, namely emojis and hashtags, on the performance of our model. We trained and evaluated our model on tweets in four different topics with heated discussions from two separate datasets, ranging from discussions about migrants to the Israel-Palestine conflict. Our multi-domain model outperforms comparable results in literature and the average results of four domain-specific models, achieving a macro-F1 score of 81.04% and 58.89% on two- and five-class tasks respectively.

Keywords: Hate Speech Detection, Contrastive Learning, Turkish Language

1. Introduction

The majority of hate speech discourse is directed towards religious, racial, gender, or ethnic groups, promoting stereotypes and negative misconceptions about these groups. According to the 2021 ADL (Anti-Defamation League) survey[1], 41% of Americans who took the survey have experienced some type of online harassment, with 35% of them reporting offensive name calling, 13% experienced stalking and 12% experienced sexual harassment.

Hateful words, images or symbols that are spread on social networks can have far reaching effects beyond strengthening stereotypes; in fact, they can provoke people to commit violent acts against the target group. In an effort to reduce the harmful effect of hate speech, there is a collective will to detect and manage (i.e. warn about, remove or counter) hate speech in social media, by governments and big companies. However, it is infeasible to manually check millions of content posted per day; therefore, research efforts have focused on hate speech detection using automated methods involving Natural Language Processing (NLP).

The first step in building detection models is to decide on a representation of the input text that is amenable to machine learning. In earlier work, traditional machine learning models were used with simple representation methods such as the TF-IDF representation (Siddiqua et al., 2019; Sevani et al., 2021; Pandey et al., 2022; Beyhan et al., 2022), while more recent approaches use deep learning with static word embeddings such as Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017), or contextualized embeddings from pre-trained language models (e.g. HateBERT (Caselli et al., 2021)).

BERT-based models using a supervised learning framework have achieved significant results in hate speech detection (Dowlagar and Mamidi, 2020; Saleh et al., 2022; Mathew et al., 2021; Beyhan et al., 2022). More recently, with the popularity of generative models, some researchers have evaluated ChatGPT's ability to detect hate speech, with the conclusion that it is not as effective as a fine-tuned the BERT model (Dehghan and Yanikoglu, 2024; Çam and Özgür, 2023).

In this study, we propose a BERT-based hate speech detection model, called BERTurk-DualCL, that uses a dual contrastive learning approach. This approach combines the benefits of contrastive learning with traditional supervised learning, to increase the quality of the learned embeddings and the robustness of the system. In addition, we evaluated the effects of paralinguistic features, namely emojis and hashtags, on system performance.

Our main contribution is contributing to the limited research on the use of dual contrastive learning, for the problems of hate speech detection and classification. A second contribution is an investigation of the effectiveness paralinguistic features (emojis and hashtags), which are widely used in tweets. Finally, we demonstrate that our multi-domain hate speech detection system results in better performance compared to the average performance obtained by models trained for individual domains.

1 https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2021

LREC-COLING 2024, pages 11745–11755
20-25 May, 2024.

The paper is organized as follows: in Section 2 and Section 3, related works and datasets for Turkish are discussed, followed by our methodology in Section 4. Finally, ablation and benchmark results are given in Section 5, followed by conclusion and future works in Section 6.

2. Related Work

We review related works in three directions: hate speech detection methods (focusing on systems developed for Turkish), contrastive learning methods, and research investigating paralinguistic features.

2.1. Hate Speech Detection Methods

Early work on hate speech detection relied on lexicons with manually selected hateful keywords (Gitari et al., 2015; Vargas et al., 2021). However, this method is often not very effective, as hate speech is not always explicit. Most of the recent works in the field of identifying hate speech for English use word embedding methods (e.g., Word2Vec, GLoVe), or large language models (e.g., BERT, RoBERTa).

There are very few works on hate speech detection for Turkish. Among the more traditional methods, Şahi et. al. (2018) used TF-IDF features to represent the stemmed input text and built a classifier model based on five machine learning algorithms, namely SVM, J48, Naive Bayes, Random Forest and Random Tree. They achieved their best results (F1 score of 70.0%) using the SVM model with the RBF kernel. Çöltekin (2020) also used the SVM classifier, but with concatenated word and character n-grams. They achieved F1 scores of 77.3% on identifying offensive tweets, 77.9% on determining whether a given offensive document is targeted, and 53.0% when classifying the targeted offensive documents into three subcategories.

More recently, Hüsünbeyi et al. (2022) developed a hybrid approach, combining deep learning models (Hierarchical Attention Network and BERT) and linguistic features. They achieved their best macro-F1 score of 90.6% and accuracy of 90.6% using with the BERT model with linguistic features.

Beyhan et al. (2022) also used a BERT base model for automatic detection and classification of hate speech. They obtained accuracies of 77.06% and 71.06% on two topics for hate speech detection; and accuracies of 72.22% and 71.74% for 5-class classification of hate speech.

Toraman et al. (2022) compared the performances of traditional approaches with deep neural models (CNN and LSTM) and transformers (BERT, RoBERTa, ConvBERT, mBERT, Megatron, and XLM-R) to detect hate speech. Their best results were obtained using ConvBERTurk and Megatron, achieving F1 scores of 78.2% and 83.0% for Turkish and English respectively.

Kurt and Demirel (2023) also used transformer models, namely DistilBERT, BERT, RoBERTa and XLM-RoBERTa. They achieved an F1 score of 82.0% and accuracy of 89.0% with XLM-RoBERTa, by multi-task learning on two large datasets, each with one million labeled samples.

Two competitions were organized recently to benchmark progress in hate speech detection in Turkish (Arın et al., 2023; Uludoğan et al., 2024). The best systems in the SIU2023-NST competition (Arın et al., 2023) achieved F1 scores of 76.87% (2-class) and 57.58% (5-class) on Anti-Refugees and Israel-Palestine conflict topics respectively. In the HSD-2Lang competition (Uludoğan et al., 2024), the best system achieved an F1 score of 69.64% in the two-class classification task on a combined test set of three topics (Anti-Refugees, Israel-Palestine conflict, and Anti-Greek sentiment in Turkey).

It is important to note that most of these works use different and/or private datasets; hence their results are not directly comparable. Similar to the later works we use a BERT model, but with reproducible results obtained on a public dataset.

2.2. Contrastive Learning Methods

Contrastive learning is a metric learning approach that aims to learn a representation space in which the representations of similar sample pairs (x, x^+) are pushed closer together, while those of dissimilar pairs (x, x^-) are pushed apart. It has achieved breakthrough performance in several computer vision tasks, such as person re-identification (Hermans et al., 2017; Khaldi and Shah, 2021), object detection (Xie et al., 2021), human-activity recognition (Tang et al., 2020), image classification (Zhang et al., 2021), and image processing (Madhusudana et al., 2022).

With success in computer vision, researchers later applied contrastive learning to various NLP tasks such as semantic textual similarity (Gao et. el., 2021; Dehghan and Amasyali, 2022, 2023), text classification (Chen et al., 2022), and hate speech detection (Lu et al., 2023). In particular, Chen et al. (2022) introduced a dual contrastive learning approach for text classification via label-aware data augmentation. Their dual contrastive model is a combination of supervised contrastive loss and cross entropy loss. Authors achieved an average accuracy of 95.43% on five English benchmark text classification datasets.

Similarly, Lu et al. (2023) introduced a BERT based dual contrastive learning approach for hate speech detection using data augmentation. Their dual contrastive model was combination of self-supervised contrastive loss, supervised contrastive

loss, and focal loss. Authors achieved an accuracy of 67.8% and a macro-F1 score of 67.2% on SemEval-2019 Task-5 hate speech dataset, which is towards against women and immigrants in English.

Similar to these two works, we use the dual contrastive learning approach, together with the cross entropy loss.

2.3. Paralinguistic Features

Feature selection has been widely studied in the context of sentiment analysis. Researchers have shown that emojis and hashtags are employed as paralinguistic features to indicate intonation and intention, more precisely than is conceivable with only punctuation characters. However, the effectiveness of emojis and hashtags has been rarely investigated for hate speech detection.

Delobelle and Berendt (2019) modified the BERT so that it can support the Unicode Standard and custom emojis, trained it on the question-answer (QA) dataset, and then were able to increase the accuracy of the base model by 5.1 points.

Liu et al. (2021) incorporated a new source of sentiment as positive, negative and neutral emojis with text. Then, they examined popular sentiment analysis algorithms including Logistic Regression, SVM, Naive Bayes classifier, Gradient Boosting Decision Tree, and a BERT-based classifier on it. They showed that emojis are effective as expanding features for improving the accuracy of sentiment analysis algorithms. Kovács et al. (2021) investigated the contribution of facial emojis to hatefulness score by measuring the Pearson correlation coefficient between scores of hatefulness/offensiveness and emojis on the OffensEval 2020 dataset. They showed that emojis were not correlated with hatefulness scores.

Corazza et al. (2020) used an LSTM model with FastText features, for hate speech detection in English, German, and Italian languages. They evaluated the impact of different features such as emoji textual description and splitting hashtags. They found that splitting hashtags could improve model performance in English and Italian languages and using emoji textual description could improve model performance slightly only in Italian language.

Li and Ning (2022) extract sentiment hashtags from tweets and convert them into a word sequence by a word segmentation tool. For example, #racismisvirus is converted to “racism is virus”. By converting hashtags into sequences of words, the semantic information is fully exposed. They compared two settings: including hashtags directly (#racismisvirus) and segmented hashtags (“racism is virus”), using the BERT model on the anti-Asian dataset. They showed that using segmented hashtags slightly improves the accuracy of the model (67.86% vs 68.85%)

Mubarak et al. (2023) categorised some common emojis into offensive, hate speech, vulgar, and violence categories. Then, they showed that Arabic tweets containing any of them are labeled as offensive. Also, by analyzing tweets, they came to the conclusion that the list of offensive emojis can be expanded: for example 65% of tweets having 💦 and 18% of tweets having a 🔨 were deemed as offensive. Additionally, the found that the top vulgar emojis are mostly used in tweets having adult content are 💋, 👅, 👠, 🐶, 🖕, 🔥 and for violence category, the most common emojis are 🔪, 👊, and 👞.

Diao et al. (2023) developed a hashtag generator model that automatically generates meaningful hashtags for incoming tweets to provide useful auxiliary signals for tweet classification. Indeed, since social media classification tasks are challenging due to the short, informal, and ambiguous nature of media posts, they increased the model’s performance by adding meaningful hashtags.

Our work provides a comprehensive evaluation about how to handle paralinguistic features for hate speech detection and supports the work that finds them useful in similar tasks.

3. Hate Speech Datasets for Turkish

In recent years, many studies have been conducted in the field of automatic detection of hate speech. However, the majority of studies in the literature target English language for which there are many resources. On the other hand, language resources for Turkish remain relatively scarce.

Şahi et. al. (2018) developed a Turkish hate speech dataset on the topic of women’s freedom , collecting tweets with the hashtag #Kıyafetimekarışma (”hands off my outfit”). The dataset contains of 1,288 tweets, with 159 instances of hate speech. Çöltekin (2020) collected an Turkish offensive speech dataset that consists of 36,232 tweets, of which approximately 19% contain some type of offensive language. It should be noted that while related, offensive speech and hate speech are seen as different.

Mayda et al. (2021) prepared a Turkish hate dataset, which included 10,224 Turkish tweets labeled hate, offensive, and none. In addition, they also assigned target labels such as ethnic, religious, sexist, and political to tweets. Hüsünbeyi et al. (2022) compiled a Turkish hate speech dataset on national and local print media news articles annotated by the Hrant Dink Foundation[2] for hate speech detection. The dataset that they obtained

2 https://hrantdink.org/en/

Table 1: Statistics for binary labelled tweets

	Ist.-Conv.	Refugees	Isr.-Pal.	Tr.-Gr.	Total
0: No HS	1086	5596	1858	694	9,234
1: Hateful	1154	1808	953	526	4,441
Total	2240	7404	2811	1220	13,675

Table 2: Statistics for multi-class labelled tweets.

	Ist.-Conv.	Refugees	Isr.-Pal.	Tr.-Gr.	Total
0: No HS	499	5596	1858	694	8647
1: Insult	380	1080	54	182	1696
2: Exclusion	118	513	248	42	921
3: Wishing harm	35	171	323	106	635
4: Threatening harm	1	44	328	196	569
Total	1033	7404	2811	1220	12468

consists of 18,316 annotated news articles published between 2016-2018, with two classes: 9,309 news articles not containing hate speech and 9,007 news articles containing hate speech.

Toraman et al. (2022) collected a dataset of hate speech in Turkish and English languages on five topics, religion, gender, racism, politics, and sports. Each dataset consists of 100k human-labeled tweets. While this is a very large dataset, it does have more label noise compared to some smaller datasets.

Beyhan et al. (2022) presented two Turkish hate speech datasets consisting of tweets collected in two separate domains, gender-based hate speech and hate speech geared towards refugees in Turkey, containing 1,206 and 1,278 samples, respectively. Arın et al. (2023) presented the SIU2023-NST dataset containing three topics: Immigrants and Refugees, Israel-Palestine Conflict and Anti-Greek Sentiment. The paper also report the results from a hate speech detection and classification contest associated with the dataset.

The last two datasets (Beyhan et al., 2022; Arın et al., 2023) are public and used in this study.

3.1. Datasets Used in This Study

We use the two publicly available datasets (Beyhan et al., 2022; Arın et al., 2023) that are partly overlapping, covering four topics: immigrants and refugees, Israel-Palestine conflict, anti-Greek sentiment and gender issues.

Each of these four topics is labelled in two ways: binary classification (based on presence or absence of hate speech) and multi-class classification (based on the type or severity of hate speech). We use 80% of data as train-split and 20% of data as test-split. The detailed statistics are given in Tables 1 and 2[3] and a brief summary about topics covered in the Istanbul-Convention and SIU2023-NST dataset

Istanbul-Convention: The Council of Europe Convention on preventing and combating violence against women is a human rights treaty signed on May 11, 2011 in Istanbul –hence known simply as Istanbul Convention (Beyhan et al., 2022). This convention was criticized by conservatives with the claim that giving too many rights to women and LGBTQ+ individuals, goes against traditional family values. With Turkey withdrawing from the convention on 20 March 2021, the controversy between conservatives and supporters increased on Twitter. This topic contains 2,240 samples in binary-class and 1,033 samples in multi-class.

Note that only a subset of the tweets were labelled with multi-class labels, and *independently* from the binary labelling; hence the number of hate-speech samples do not match exactly.

Immigrants and Refugees in Turkey: In recent years, due to the civil wars in Syria and Afghanistan, countless immigrants and refugees from these countries have found refuge in Turkey. According to the latest statistics, there are about 3.7 million Syrians and about 300,000 Afghans who settled in Turkey. While public opinion was welcoming at the beginning of the refugee crisis, the problems caused by the large number of asylum seekers and the widespread misconception that refugees can

3 The discrepancy in the total number of these two tables is due to the fact that the labeling was done separately for binary and multi class cases.

Table 3: Sample emojis with their textual aliases in Turkish

Emoji	Textual Alias		Emoji	Textual Alias	
	En	Tr		En	Tr
😀	Grinning face	Sırıtan yüz	😭	Loudly crying face	Yüksek sesle ağlayan yüz
😉	Winking face	Göz kırpan yüz	😑	Expressionless face	İfadesiz yüz
🖤	Black heart	Siyah kalp	😈	Angry face with horns	Boynuzlu kızgın yüz
🔥	Fire	Ateş	😱	Face screaming in fear	korku içinde çığlık atan yüz
😔	Pensive face	Dalgın yüz	👊	Oncoming fist	Yaklaşan yumruk
🚫	Prohibited	Yasak	💘	Heart with arrow	Oklu kalp
😠	Angry face	kızgın yüz	😢	Crying face	Ağlayan yüz
😕	Confused face	Şaşkın yüz	🏴	Black flag	Siyah bayrak
💥	Collision	Çarpışma	😕	Slightly frowning face	Hafifçe kaşlarını çatan yüz
😠	Angry face	kızgın yüz	😖	Confounded face	Kafası çok karışık yüz
😏	Smirking face	Sırıtan yüz	😥	Sad but relieved face	Üzgün ama rahatlamış yüz

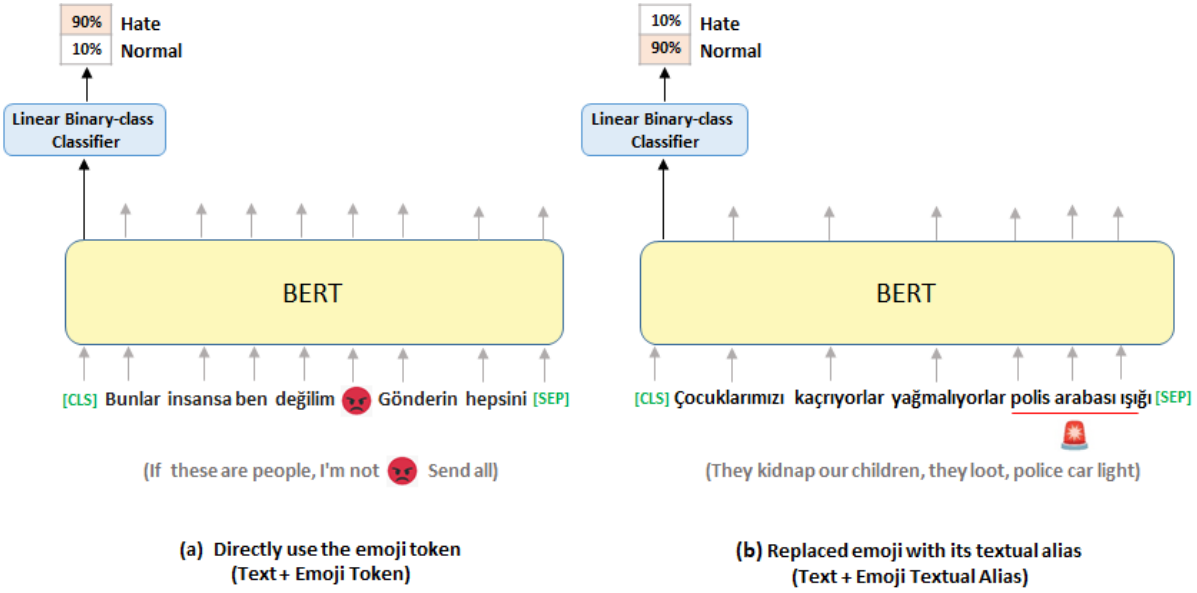

Figure 1: Two alternative tokenization of emojis are evaluated in this study.

be given rights that the Turkish people do not have, have increased negative feelings towards them. Hence, this caused an increase in hate speech towards them on social networks. This topic contains 7,404 samples in both binary-class and multi-class.

Israel-Palestine Conflict: The Israel-Palestine conflict is one of the most enduring conflicts in the world, starting in the mid-20th century. Despite long-term efforts for peace and general reconciliation, the problem unfortunately persists very much, with pro-Israeli and pro-Palestinians deeply differing in opinions. With ongoing and increasing conflict turning into war at times, this is a hot topic of discussion in Turkey. This topic contains 2,811 samples in both binary-class and multi-class.

Anti-Greek Sentiment in Turkey: Anti-Hellenism or Hellenophobia (simply known as Anti-Greek) refers to hatred and prejudice against Greeks, the Hellenic Republic, and the Greek culture. Since the treaty of Lausanne, Turkey and Greece have been at odds over the sovereignty of the Aegean islands, territorial waters, flight zones, and the violation of the rights of their respective minorities. In particular, in the summer of 2022, the Greeks began to increase their military presence on the islands, and this action intensified the rhetoric of politicians from Ankara and Athens, as well as the two populations, against each other as Turkish elections approached. This topic contains 1,220 samples in both binary-class and multi-class. Further details of the dataset is given in (Beyhan et al., 2022; Arın et al., 2023).

Statistics for each sub-domain of our dataset for binary and multi class problems are given in Tables 1 and 2.

4. Methodology

In this section, we first introduce various paralingusitic features that we consider in our hate speech detection model (Section 4.1). Then, we introduce our dual contrastive learning model for hate speech detection using the selected features (Section 4.2).

4.1. Paralingusitic Feature Selection

Detecting hate speech on social media is a difficult task because social media posts are usually informal and include mentions, URLs, and paralinguistic features (e.g., emojis and hashtags). Therefore, preprocessing of the text data is commonly done to remove some of these elements and reduce the linguistic variance. Since URLs and usernames do not generally provide useful information towards hate speech detection or classification, we remove them, but we studied the effect of other paralinguistic features –namely emojis and hashtags– on the performance of hate speech detection.

Emojis are Unicode graphic symbols that are used as abbreviations for thoughts and emotions. Graphic emojis have become an integral part of today's conversations, so that a thumbs-up/thumbs-down emoji can indicate the speaker's agreement or disagreement without any words.

Similarly, hashtags are very important in social media because they allow to link messages around a particular hashtag. However, hashtags are often removed during preprocessing in an effort to simplify the subsequent modeling (Bhatnagar and Choubey, 2021; De Arriba et al., 2021), but hashtags are sometimes used as words in the middle of a sentence and removing them destroys the meaning of the whole sentence.

To study affect of emojis and hashtags on model performance, we considered five feature subsets:

1. **Text:** We removed all URLs, mentions, emojis and hashtags.

2. **Text + Emoji Token:** We directly added the emoji tokens to the BERTurk tokenizer[4] (an uncased BERT model for Turkish). We found that the BERTurk tokenizer only supports 120 out of the 4,733 emojis present in the commonly used Emoji Library [5], which amounts to about 2.5%. For instance, among the sentiment indicating emojis, the 😊 emojis are covered but the 😠 emoji was not. To improve the coverage, we added the most frequent 185 emojis to BERTurk tokenizer and increased the coverage to 6.4%.

3. **Text + Emoji Textual Alias:** We replaced emojis with with their textual aliases and append it to the end of the input text. As the Emoji Library does not support Turkish language, we created our own dictionary of emoji textual aliases in Turkish language. Figure 1 and Table 3 show tokenizing process and some of the emoji tokens and its aliases that we used for options (2) and (3), respectively. In this option, we also tested the effect of removing duplicate emojis.

4. **Text + Hashtag:** We just removed # mark and the hashtag text remains unchanged. We also did not split the hashtag.

5. **Text + Emoji Token + Hashtag:** This option is a combination of options (2) and (4).

4.2. Dual Contrastive Learning Model

We design our model using transfer learning, with a single layer on top of the learned encoder (BERT) to predict the hate speech categories. As we use a contrastive design, the input of our model is a batch of tweets including both hate and non-hate text (binary-class or multi-class). Figure 2 shows the framework.

The model is trained using a dual contrastive loss function that is a combination of cross-entropy loss (L_{CE}) and supervised contrastive loss (L_{SCL}):

$$L_{DualCL} = L_{CE} + \lambda L_{SCL} \quad (1)$$

where $\lambda \in [0, 1]$ is a weighting hyperparameter that controls the impact of these two loss functions. We tried different λ values and observed that the best performance for $\lambda = 0.5$. The cross-entropy loss for classification is defined as:

$$L_{CE} = -\sum_{i=1}^{N} y_i log(\hat{y}_i) \quad (2)$$

where y_i is the target value for the ith input and $\hat{y}_i$ is the prediction. The two objectives learn classifier parameters and improve the the quality of the representations of the features, simultaneously.

As our supervised contrastive loss (L_{SCL}), we use the loss introduced in (Khosla et al., 2021). They extended the self-supervised NT-Xent loss (Chen et al., 2020) to the fully supervised setting by adding multiple positives from the same class. The NT-Xent loss only accepts one positive sample x_i^+, obtained via augmentation, for an anchor x_i, and uses the other samples in the mini-batch as negatives, obtaining a mini-batch of $(x_i, x_i^+, x_1^-, ..., x_K^-)$. Supervised contrastive loss extends the number of positives in the mini-batch to P positive samples from the training set, obtaining a mini-batch of

4 https://huggingface.co/dbmdz/bert-base-turkish-uncased

5 https://carpedm20.github.io/emoji/docs/index.html

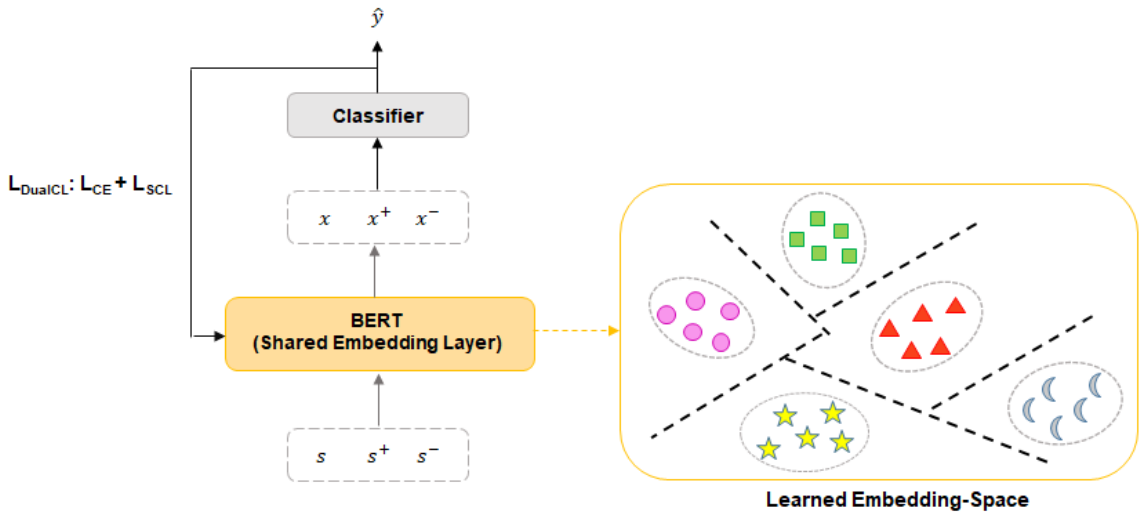

Figure 2: Our Dual Contrastive Learning framework.

$(x_i, x_1^+, ..., x_P^+, x_1^-, ..., x_K^-)$. Our SCL is defined as:

$$L_{SCL} = -\frac{1}{P}\sum_{p=1}^{P} \log \frac{e^{(x_i \cdot x_p^+/\tau)}}{\sum_{k=1}^{K} e^{(x_i \cdot x_k^-/\tau)}} \quad (3)$$

where x_i is the embedding vector of sentence s_i, x_p^+ and x_k^- indicate positive and negative samples respectively; the symbol $\cdot$ denotes the inner product and $\tau \in R^+$ denotes the temperature parameter set to 0.1.

5. Experiments

We have evaluated the effectiveness of various feature sets mentioned in Section sec:feature-set first, with results given in Table 4. Then using the best feature set, we evaluate the effectiveness of the proposed approach by fine-tuning the classifier model using only the cross-entropy loss or the DualCL loss, with results given in Table 5.

5.1. Training Setting

We use a pretrained BERT transformer with a linear classification head on top of the pooled output, using Huggingface Transformer package. We start fine-tuning from pretrained checkpoint of BERTurk.

For the so called CE method (see Table 5) where we only use the cross-entropy loss for comparison, we conducted experiments for binary-class and multi-class classification using stratified 10-fold cross validation. The sample size of train, validation and test sets were 70%, 10% and 20%, respectively for each run. We trained the models for 10 epochs with batch size of 16.

For the DualCL method, we trained our model for 4 epochs with a batch size of 16 for binary-class and multi-class classification tasks, separately. We used an NVIDIA A100 GPU with CUDA 11.8. As our training dataset is imbalanced, we use both accuracy and macro-F1 score as evaluation metrics.

5.2. Results

Feature selection results: The result of selecting various features on base model performance are given in Table 4. We first observe that paralinguistic features carry important information about the meaning and affect classification performance. Hashtags are particularly useful and bring improvements in both the macro-F1 score (79.68% vs 77.04%) and accuracy (83.74% vs 81.58%) over the baseline model of using only text features.

Including emoji token or emoji textual alias also improves the performance, but they are not found as valuable as hashtags.

Binary and multi-class classification results: We evaluated four different feature options that performed best in the previous experiment and evaluated the proposed dual contrastive loss, for both classification problems.

As seen in Table 5, BERTurk model trained with the DualCL loss (BERTurk-DualCL) showed better performance compared to the baseline model that only uses the cross-entropy loss (BERTurk-CE model), for all feature sets. This shows that the proposed dual contrastive learning is useful for this problem, as was shown for other NLP problems before (Gao et. el., 2021; Chen et al., 2022; Dehghan and Amasyali, 2022, 2023).

The best performance is obtained with the ”Text + Hashtag + Emoji Token” feature set for BERTurk-DualCL, with 81.04% macro-F1 and 84.62% accuracy on the detection problem (2-class classification). The best 5-class classification results were also obtained with this model and feature set.

Table 4: Feature selection results using 10-fold cross-validation with the base model, using the cross-entropy (CE) loss. Best results are shown in bold.

	Feature Set	Multi-domain (2-class)	
		Macro-F1	Acc.
BERTurk	Text	77.04	81.58
	Text + Emoji Token	77.13	82.17
	Text + Emoji Textual Alias	77.53	82.28
	Text + Emoji Textual Alias (remove duplicates)	77.73	82.57
	Text + Hashtag	79.68	83.74
	Text + Hashtag + Emoji Token	79.49	**83.81**
	Text + Hashtag + Emoji Textual Alias	**79.69**	83.77
	Text + Hashtag + Emoji Textual Alias (remove duplicates)	79.68	83.59

Table 5: Multi-domain results of BERTurk-DualCL, in comparison to the proposed model with baselines. Best results are in bold.

Model	Features	Method	Multi-domain (2-class)		Multi-domain (5-class)	
			Macro-F1	Acc.	Macro-F1	Acc.
BERTurk	Text + Hashtag	CE	79.68	83.74	56.35	80.48
		DualCL	80.36	84.53	58.02	81.04
	Text + Hashtag + Emoji Token	CE	79.49	83.81	54.99	80.56
		DualCL	**81.04**	**84.62**	**58.89**	**81.52**
	Text + Hashtag + Emoji Textual Alias	CE	79.69	83.77	54.67	80.93
		DualCL	80.30	84.61	56.88	80.08
	Text + Hashtag + Emoji Textual Alias (remove duplicates)	CE	79.68	83.59	54.14	80.64
		DualCL	80.19	84.21	57.73	81.12

Confusion matrices of BERTurk-DualCL model for binary and multi-class classification are shown in Figure 3.

Comparison with single-domain hate speech detection: Table 6 gives the results for the BERTurk-DualCL model trained for individual topics, using the "Text + Hashtag + Emoji Token" feature set. As seen in this table, training the model in multiple domains achieves higher accuracies (84.62% and 81.52% for 2-class and 5-class problems respectively, see Table 5) compared to the average of individually trained models (84.57% and 78.15%, see Table 6). We think that the larger improvement for the 5-class problem is due to the low number of samples in some classes in individual domains.

Comparison to the literature: Our results are only directly comparable to those obtained on the 5-class problem for the Israel-Palestine topic in the SIU2023-NST competition (Arın et al., 2023), which is 57.58% macro-F1 compared to the 58.89% obtained in this work. The other relevant result from the SIU2023-NST competition is 76.87% macro-F1 for the 2-class classification for a subset of the Refugee dataset (5,854 tweets as opposed to the 7,404 used in this work[6]).

While not directly comparable, the best results in the HSD-2Lang competition (Uludoğan et al., 2024) was 69.64% macro-F1 in the two-class classification task on a combined test set of three topics (Anti-Refugees, Israel-Palestine conflict, and Anti-Greek sentiment in Turkey), which is lower than the 81.04% obtained in this work for two-class classification on four topics.

As for the results in (Beyhan et al., 2022), they are obtained on a 1206-sample subset of Ist.-Conv. topic and 1278-sample subset of Refugees topic, with 77.06% and 71.06% accuracy respectively.

6. Conclusions and Future Works

We present a Turkish hate speech detection and classification model that is trained using the dual contrastive loss which is a combination of cross-entropy loss and supervised contrastive loss. We also evaluate the effectiveness of paralinguistic

6 Tweets less than 100 characters were not included in SIU2023-NST competition dataset.

Table 6: Single-domain results of BERTurk-DualCL. Four domain-specific models are trained and tested only on single topic each, using the BERTurk model trained with "Text + Hashtag + Emoji Token" features and DualCL loss. Average performance is shown in bold.

	Istanbul Conv.		Refugees		Israel-Palestine		Tr.-Gr.		Average	
	Single-domain (2-class)									
	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.	M-F1	Acc.
BERTurk-DualCL	79.35	79.46	74.86	82.43	80.51	90.36	85.40	86.06	**80.03**	**84.57**
Single-domain (5-class)										
BERTurk-DualCL	42.47	69.41	40.12	77.90	48.10	89.49	51.69	75.81	**45.59**	**78.15**

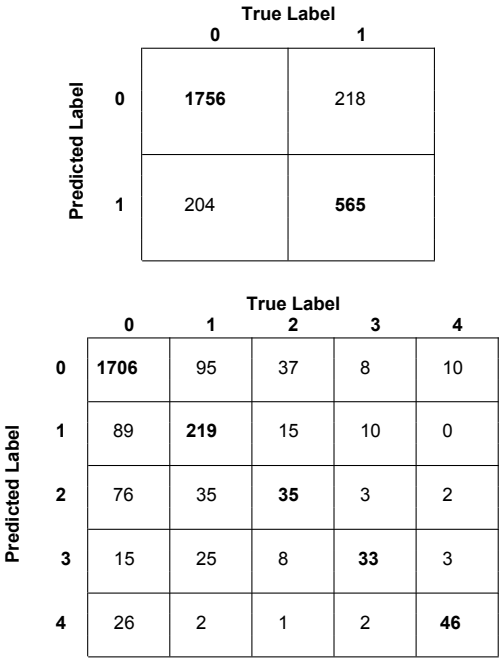

Figure 3: Confusion matrix of multi-domain BERTurk-DualCL model for binary (top) and multi-class (bottom) classification on the test dataset.

features, and use the best feature subset (text, emoji tokens and hashtags) in the final model.

We evaluated our model on a combined hate speech dataset covering four topics: Istanbul-Convention, Immigrants and Refugees in Turkey, Israel-Palestine Conflict, and Anti-Greek sentiment in Turkey. Experimental results show a significant improvement compared to the baseline model which is trained with only the cross-entropy loss and without the paralingusitic features. Our findings show that hashtags and emojis are important features in hate speech detection in tweets.

Unlike other hate detection models, which are often trained on one or two topics, we were able to efficiently train a single multi-domain model which obtained better performance than the average of four models that were each trained on a single topic.

As future studies, we plan to try different data augmentation techniques to increase the performance of our hate speech detection model. In addition, we will also work to collect a hate speech lexicon and hate speech hashtags in Turkish language.

7. Acknowledgements

This work was supported by the EU project "Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity" (EuropeAid/170389/DD/ACT/Multi), carried out by the Hrant Dink Foundation.

8. Bibliographical References

Arın, I., Işık, Z., Kutal, S., Dehghan, S., Özgür, A., Yanıkoğlu, B. (2023). SIU2023-NST - Hate Speech Detection Contest. 31. IEEE Conference on Signal Processing and Communications Applications, Istanbul.

Beyhan F., Çarık B., Arın, I., Terzioğlu, A., Yanıkoğlu, B., Yeniterzi, R. (2022). A Turkish Hate Speech Dataset and Detection System. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pp. 4177–4185.

Bhatnagar, S., Choubey, N. (2021). Making sense of tweets using sentiment analysis on closely related topics. Social Network Analysis and Mining, 11, 44.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with sub word information. Trans. Assoc. Comput. Linguist., pp. 135-146.

Caselli, T., Basile, V., Mitrovic, J., Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. Proceedings of

the 5th Workshop on Online Abuse and Harms (WOAH 2021), pp. 17–25.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Çam, N. B., Ozgur, A. (2023). Evaluation of ChatGPT and BERT-based models for Turkish hate speech detection. In Proceedings of the International Conference on Computer Science and Engineering (UBMK).

Çöltekin, Ç. (2020). A corpus of turkish offensive language on social media. Proceedings of the 12th Language Resources and Evaluation Conference, pp. 6174–6184.

Chen, Q., Zhang, R., Zheng, Y., Mao, Y. (2022). Dual Contrastive Learning: Text Classification via Label-Aware Data Augmentation. arXiv:2201.08702.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., Villata, S. (2020). A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT), 20(2), 1-22.

De Arriba, A., Oriol, M., Franch, X. (2021). Applying sentiment analysis on Spanish tweets using BETO. A: Iberian Languages Evaluation Forum. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), pp. 1-8.

Dehghan, S., Amasyali, M.F. (2022). SupMPN: Supervised Multiple Positives and Negatives Contrastive Learning Model for Semantic Textual Similarity. Applied Sciences, 12:9659.

Dehghan, S., Amasyali, M.F. (2023). SelfCCL: Curriculum Contrastive Learning by Transferring Self-Taught Knowledge for Fine-Tuning BERT. Applied Sciences, Vol. 13(3):1913.

Dehghan, S., Yanikoglu, B. (2024). Evaluating ChatGPT's Ability to Detect Hate Speech in Turkish Tweets. In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 54–59, St. Julians, Malta. Association for Computational Linguistics.

Delobelle, P., Berendt, B. (2019). Time to Take Emoji Seriously: They Vastly Improve Casual Conversational Models. arXiv:1910.13793.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 4171-4186.

Diao, S., Keh, S. S., Pan, L., Tian, Z., Song, Y., Zhang, T. (2023). Hashtag-Guided Low-Resource Tweet Classification. arXiv:2302.10143.

Dowlagar, S., Mamidi, R. (2020). HASOCOne FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection. FIRE '20, Forum for Information Retrieval Evaluation, Hyderabad, India.

Gao, T., Yao, X., Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic.

Gitari, N. D., Zuping, Z., Damien, H., Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. International Journal of Multimedia and Ubiquitous Engineering Vol.10, No.4, pp.215-230.

Hermans, A., Beyer, L., Leibe, B. (2017). In defence of triplet loss for person re-identification. arXiv:1703.07737.

Hüsünbeyi, Z. M., Akar, D., Özgür, A. (2022). Identifying Hate Speech Using Neural Networks and Discourse Analysis Techniques. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pp. 32–41.

Khaldi, K., Shah, S. (2021). CUPR: Contrastive Unsupervised Learning for Person Re-identification. In Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021), Volume 5, pp. 92-100.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D. (2021). Supervised Contrastive Learning. arXiv:2004.11362.

Kovács, G., Alonso, P., Saini, R. (2021). Challenges of Hate Speech Detection in Social Media. SN COMPUT. SCI. 2, 95.

Kurt, M. S., Yücel Demirel, E. (2023). Türkçe Hakaret ve Nefret Söylemi Otomatik Tespit Modeli. Veri Bilimi, Vol. 6, No. 1, pp. 61-73.

Li, J., Ning, Y. (2022). Anti-Asian Hate Speech Detection via Data Augmented Semantic Relation Inference. arXiv:2204.07010.

Liu, C., Fang, F., Lin X., Cai, T., Tan, X., Liu, J., Lu, X. (2021). Improving sentiment analysis accuracy with emoji embedding. Journal of Safety Science and Resilience, Volume 2, Issue 4, pp. 246-252.

Lu, J., Lin, H., Zhang, X., Li, Z., Zhang, T., Zong, L., Ma, F., Xu, B. (2023). Hate Speech Detection via Dual Contrastive Learning. arXiv:2307.05578.

Maaten, L.V. D., Hinton, G. E. (2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research, 9, pp. 2579–2605.

Madhusudana, P.C., Birkbeck, N., Wang, Y., Adsumilli, B., Bovik, A.C. (2022). Image Quality Assessment using Contrastive Learning. IEEE Transactions on Image Processing. 31, pp. 4149-4161.

Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv:2012.10289.

Mayda, I., Demir, Y. E., Dalyan, T., Diri, B. (2021). Hate Speech Dataset from Turkish Tweets. 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 2021, pp. 1-6.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mubarak, H., Hassan, S., Chowdhury S. A. (2023). Emojis as anchors to detect Arabic offensive language and hate speech. Natural Language Engineering. 2023; 29(6):1436-1457.

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar.

Pandey, Y., Sharma, M., Siddiqui, M.K., Yadav, S.S. (2022). Hate Speech Detection Model Using Bag of Words and Naïve Bayes. In: Tiwari, S., Trivedi, M.C., Kolhe, M.L., Mishra, K., Singh, B.K. (eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 318. Springer, Singapore.

Saleh, H., Alhothali, A., Moria, K. (2022). Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model. arXiv:2111.01515.

Sevani, N., Soenandi, I.A., Wijaya, J. (2021). Detection of Hate Speech by Employing Support Vector Machine with Word2Vec Model. 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), Malang, Indonesia, pp. 1-5.

Siddiqua, U. A., Chy, A. N., Aono, M. (2019). KDE-HatEval at SemEval-2019 Task 5: A Neural Network Model for Detecting Hate Speech in Twitter. Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA. pp. 365-370.

Şahi, H., Kılıç, Y., Sağlam, R.B. (2018). Automated detection of hate speech towards woman on twitter. International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, pp. 533-536.

Tang, C. I., Perez-Pozuelo, I., Spathis, D., Mascolo, C. (2020). Exploring Contrastive Learning in Human Activity Recognition for Healthcare. Presented at the Machine Learning for Mobile Health Workshop at NeurIPS 2020, Vancouver, BC, Canada.

Toraman, C., Şahinuç, F., Yilmaz, E. (2022). Large-Scale Hate Speech Detection with Cross-Domain Transfer. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, pp. 2215–2225.

Uludoğan, G., Dehghan, S., Arin, I., Erol, E., Yanikoglu, B., Özgür, A. (2024). Overview of the Hate Speech Detection in Turkish and Arabic Tweets (HSD-2Lang) Shared Task at CASE 2024. In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), pages 229–233, St. Julians, Malta. Association for Computational Linguistics.

Vargas, F., de Góes, F. R., Carvalho, I., Benevenuto, F., Pardo, T. (2021). Contextual-Lexicon Approach for Abusive Language Detection. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp. 1438–1447.

Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P. (2021). DetCo: Unsupervised Contrastive Learning for Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 8372-8381.

Zhang, Z., Jang, J., Trabelsi, C., Li, R., Sanner, S., Jeoong, Y., Shim, D. (2021). ExCon: Explanation-driven Supervised Contrastive Learning for Image Classification. arXiv:2111.14271.