Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4177–4185
Marseille, 20-25 June 2022

A Turkish Hate Speech Dataset and Detection System

Fatih Beyhan[1,2], Buse Çarık[1,2], İnanç Arın[1,2], Ayşecan Terzioğlu[3], Berrin Yanıkoğlu[1,2], Reyyan Yeniterzi[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
[3]Faculty of Arts and Social Sciences, Sabanci University, Istanbul, Turkey 34956
{fatihbeyhan,busecarik,inanc,aysecan,berrin,reyyan}@sabanciuniv.edu

Abstract

Social media posts containing hate speech are reproduced and redistributed at an accelerated pace, reaching greater audiences at a higher speed. We present a machine learning system for automatic detection of hate speech in Turkish, along with a hate speech dataset consisting of tweets collected in two separate domains. We first adopted a definition for hate speech that is in line with our goals and amenable to easy annotation; then designed the annotation schema for annotating the collected tweets. The Istanbul Convention dataset consists of tweets posted following the withdrawal of Turkey from the Istanbul Convention. The Refugees dataset was created by collecting tweets about immigrants by filtering based on commonly used keywords related to immigrants. Finally, we have developed a hate speech detection system using the transformer architecture (BERTurk), to be used as a baseline for the collected dataset. The binary classification accuracy is 77% when the system is evaluated using 5-fold cross validation on the Istanbul Convention dataset and 71% for the Refugee dataset. We also tested a regression model with 0.66 and 0.83 RMSE on a scale of [0-4], for the Istanbul Convention and Refugees datasets.

Keywords: Hate speech detection, Deep learning, Turkish

1. Introduction

Definition. Hate speech refers to discourse that targets a specific group based on race, gender, religion, sexual orientation, etc., and indicates some level of hatred towards them. In fact, there is no fully agreed convention on what constitutes hate speech (Davidson et al., 2017; Poletto et al., 2017), however, there have been various attempts to define it. In a commonly accepted definition, hate speech is described as "speech that targets disadvantaged social groups in a manner that is potentially harmful to them" (Jacobs et al., 1998; Walker, 1994).

Hate speech comes in a *spectrum*, ranging from mild stereotyping to open calls to violence toward the target group, partially explaining the difficulty in coming up with a description and annotating collected samples. Some works in the literature use a binary categorization (hate speech or not), while others use a finer level of categorization (Davidson et al., 2017; Poletto et al., 2017). Furthermore, what constitutes hate speech is subjective to some level and may require background knowledge on the topic to detect.

Aim. Negative discourses including hate speech are reproduced and redistributed on social media and internet at an accelerated pace, allowing access at a higher speed and to greater audiences in number. Online hate speech does not only take place in unofficial or anonymous accounts but is also observed in institutional and officially recognized discourses. The rising threat of online hate speech has been recognized and received internationally organized responses, such as UNESCO actions and researchers sharing as hate speech datasets to boost automatic methods (Basile et al., 2019; Gagliardone et al., 2015).

The aim of this project is to detect and evaluate hate speech in the tweets about the prevalent public issues in Turkey. Being a politically polarized and socially dynamic country, there have been many different public issues considered for this research. However, we have decided to focus on two topics, Istanbul Convention and attitudes against the refugees in Turkey, because of the high volume and visibility of these two topics on Twitter. In doing so, we aim to capture different aspects of hate speech, such as gender-based hate speech and hate speech based on ethnicity, race, or nationality through an analysis of how the hate-speech dominates the two most pertinent public discussions on the Twitter.

Gender-based hate speech. One dataset collected in this work is related to gender. The Istanbul Convention[1] is issued by the Europe's leading human rights organization, the Council of Europe, with the aim of establishing "standards on preventing, protecting against and prosecuting the most severe and widespread forms of gender-based violence across Europe". While the Istanbul Convention is celebrated among the progressive politicians, feminist and LGBTI+ activists, organizations, and institutions all across Europe, the conservative governments, institutions, and politicians criticized the convention, arguing that it harms the "traditional family values" by giving too many rights to women and LGBTI+ (Burnett, 2021). The criticisms caused massive disinformation campaigns, wrongly claiming that the Istanbul Convention promotes gay marriage and would degrade the family union. The two opposing sides have often clashed in the social media, especially on Twitter in Turkey, in accord with the political and social polar-

[1]The full name is "The Convention on Preventing and Combating Violence Against Women and Domestic Violence", but it is commonly referred to as the Istanbul Convention, since the European ministers had signed the document, prepared in that convention in Istanbul, in 2011.

ization between conservative and progressive people in the society. The tension has considerably increased in these clashes when Turkey became the first country, which withdrew from the Istanbul Convention in March, 2021. The tension between both sides led to the increase in the use of hate speech and cyberbullying, similar to what happened in Bulgaria, which (Bankov, 2020) explains with the "e-crowd effect", where emotions and opinions dominate rationality and facts.

Hate speech geared towards refugees. Hate speech against the refugees increased since the 2010s as a high number of Syrians escaped from the war in their country to Turkey. Today, according to the Turkey's Directorate of Migration Management there are 3.7 million registered Syrians in Turkey (ANSA, 2021). In addition, the Afghans who are escaping from the Taliban regime often end up settling down in Turkey, and their official number is around 300.000 (Sanderson, 2021). The Turkish state tries to meet the basic needs of these refugees, such as health care, housing, and employment, using the 4.3 million Euros donated to the Turkish government by the European Union for this purpose (Memisoglu and Ilgit, 2017). However, this help often remains inadequate given the high number of refugees, and the problems of coordination among the state, municipalities, and non-governmental organizations (NGOs). While the common sentiment in the beginning of the refugee crisis was more welcoming, the problems due to the very large number of refugees and the common misconceptions that the refugees may be given rights that are not available for the Turkish population led to an increase in negative sentiment against the refugees. The COVID pandemic and the problems in managing the pandemic, current economic crises, and political polarization in Turkey, between the people pro- or against the current government, also exacerbate these misconceptions and hate speech against the refugees in the society, and this is directly reflected on Twitter.

Organization. In the remainder of this paper, we discuss related work (Section 2.), focusing especially on annotated datasets and automated methods for detecting hate speech. In Section 3., we provide our definition of hate speech and the annotation categories and setting. In Section 4., we describe the hate speech dataset collected in the scope of this work on the above two topics. In Section 5., we present a system to automatically detect hate speech and provide benchmark results on the collected dataset.

Our contributions are as follows:

- We collected a dataset of 1206 tweets related to gender and sexual orientation based violence ("Istanbul Convention dataset") and 1278 tweets related to refugees in Turkey ("Refugee dataset"). After removing tweets with conflicting category annotations, there are 1033 and 1278 tweets, respectively. The dataset is publicly available at `https://github.com/verimsu/Turkish-HS-Dataset`.

- Hate speech definition and annotation categories have been considered carefully and in several iterations, to be amenable to easy and unambiguous annotation.

- We developed a hate speech detection system using the transformer architecture in both classification and regression settings, to be used as a baseline for the collected dataset. Our results for 5-fold cross validation are as follows. We have obtained an accuracy of near 80% and 72% on the Istanbul Convention dataset, on the binary (hate speech or not) and 5-class classification problems (different levels of hate speech), respectively. The results for the Refugee dataset are approximately 71% in both binary and 5-class settings. The regression model's cross validation score for the Istanbul Convention dataset is 0.66 RMSE, while the RMSE score for the Refugee dataset is 0.83.

2. Related Work

Unfortunately, it is quite common to reproduce and distribute negative discourses on individuals and groups based on gender, race, ethnicity, religion, and political ideologies through social media. Even some international organizations like Hatebase (Hatebase.org, 2022) and UNESCO (Gagliardone et al., 2015) mention the rising threat of hate speech through their communication channels. Consequently, there have been many studies in social sciences that study and discuss the concept of hate speech (Karaman and Işıklı, 2016; Dondurucu and Uluçay, 2015; Hüseyin and Öksüz, 2020). In the scope of this work, we list studies where the definition of hate speech or computational approaches to detecting hate speech are the focus.

In one of the earlier works, Waseem and Hovy (2016) focused on 2 branches of hate speech; racism, and sexism. They used several rules to decide on whether a tweet has hate speech. Authors achieved their best F1 score (0.739) using character n-grams of lengths up to 4, along with gender as an additional feature.

Davidson et al. (2017) focused on differentiating hate speech from offensive language. In their definition of hate speech, they stated that for the existence of hate speech, a group must have been targeted. They defined 3 classes (Hate, Offensive, and Neither); and their best performing model had an overall F1 score of 0.90.

Similarly, Poletto et al. (2017) aimed to distinguish between hate speech, offensive language, and neither, in a study focusing on hate speech towards refugees and Muslims in Italy. The authors used keywords for groups that are the target of hate speech in order to distinguish between offensive language and hate speech and mentioned why lexical detection methods failed to distinguish these two concepts from each other. The keywords are geared towards identifying prejudice against the groups that are the target of hate speech. The category with the highest agreement was the presence of hate speech with Cohen's kappa coefficient of 0.54.

Ross et al. (2017) worked on German tweets, and annotated tweets in terms of hate speech and offensive language; however, instead of treating hate speech and offensive language as disjoint concepts, they labeled hate speech as binary and indicated scales for offensive language. Their findings show that individuals tend to interpret the meaning of the hate speech aligned with their own opinions, and

No Hate speech	Insult	Exclusion	Wishing Harm	Threatening Harm
Do not give priority to these people for vaccines.	Do you see how these animals are jumping over the border?	I do not want my kids to be in the same classroom as Syrian kids.	I hope these refugees drown with those boats and cannot come back!	I am so full that I will kill the first Syrian that comes my way!

Table 1: The 5-level hate speech categories used in annotation, along with samples given in the tool to guide the annotators.

this leads to an unreliable labeling process in general. They added that it would be better to have a system which has more options than just having binary yes/no labels for a hate speech detection problem. They also believe that more instructions and guidelines should be provided for the individual annotators as well.

In (Poletto et al., 2017; Sanguinetti et al., 2018), after deciding on their target groups (immigrants), authors filtered their dataset with stereotypical keyword groups. They labelled their dataset for hate speech and offensive language with ordinal labels, using a hate speech definition based on whether the tweet has one of their *target groups* and there is an *action* against the group. In their work, they also mentioned how important and challenging it is to have an agreement among different annotators due to the individual biases. The best Cohen kappa coefficient was 0.48 even if the data was annotated by an expert team. They concluded that further refinements should be applied for such a challenging problem.

Mathew et al. (2020) adopt the annotation format of (Davidson et al., 2017) and use the same labels; however, they also annotated the dataset in two different perspectives: target community and the rationales. Target communities are predefined groups that consist of races, religions, etc. The rationales are the spans of the text where their labelling decisions are based. The best F1 score they obtained by using BERT was 0.698.

Çöltekin (2020) has developed a Turkish offensive speech dataset that consists of randomly sampled micro-blog posts from Twitter. This work is the most similar to ours, but offensive speech and hate speech are distinct, as discussed in numerous studies (Davidson et al., 2017; Poletto et al., 2017). The collected offensive speech corpus contains 36.232 tweets sampled randomly from the Twitter stream during a period of 18 months between April 2018 to September 2019. They have trained three separate classifiers: a binary classifier discriminating offensive tweets from non-offensive tweets; another binary classifier that predicts whether an offensive tweet is targeted or not; and finally, a three-way classifier that predicts the target type (individual, group, or other) of a targeted offensive tweet. They report F1 scores 0.773, 0.779, and 0.53 respectively for each classifier and found a clear elevation of offensive language use, particularly offensive posts with a group target, during two elections within the time span of their data.

3. Hate Speech Definition and Annotation

As there is no fully agreed convention on what constitutes hate speech (Davidson et al., 2017; Poletto et al., 2017), we first worked on a definition in Section 3.1.. Later on, we worked on the levels of annotation for hate speech categorization in Section 3.2.

3.1. Hate Speech Definition

Hate speech is described as "speech that targets disadvantaged social groups in a manner that is potentially harmful to them" (Jacobs et al., 1998; Walker, 1994). While focusing on disadvantaged social groups (based on ethnicity, religion, gender, etc), we expanded the action definition to be more in line with that of (Davidson et al., 2017) as *"language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group"*. Our motivation to concentrate on the disadvantaged groups is based on the large number of hateful comments on social media, which would result in the over-use of the term hate speech.

3.2. Annotation Levels

Deciding on the number of levels for hate speech annotation was done as an iterative process, in two main stages. We first started annotating the Istanbul Convention dataset on a scale of 1-10. However, annotators indicated that they had difficulties in selecting the level. We then switched to a 5-level annotation with clear category titles and examples shown on the annotation tool. The initial set of annotated tweets before the switch are excluded from the dataset.

The hate speech spectrum is divided into four categories, as indicated in Table 1: 1) derogatory comments ("they are worthless"); 2) exclusion comments ("they should go home"); 3) harm-wishing comments ("I hope they die"); and 4) harm-threats ("I would kill ..."). We have found that this 5-level scheme has been conceived as easier and less ambiguous by the annotators.

This categorization is in line with the study in the Italian language, which scales the intensity of hate speech on a range of 1 to 4 (Sanguinetti et al., 2018). They describe each level of intensity as follows; insult based on the minority group as level 1, statements that ignore the fundamental rights of these people as level 2, wishing them to be subjected to violence by others as level 3, and openly threatening and calling for violent actions towards them as level 4 (Sanguinetti et al., 2018).

The annotators were each given a tutorial where they learned about our hate speech definition; the concept of target groups; and that offensive language (no precise definition was given) and hate speech were different. While our definition requires hate speech to contain a disadvantaged group, the annotators were told to label hate speech without the target as well, so as to leave this decision to researchers using the dataset.

3.3. Annotation Tool

During the annotation process we tried to retrieve as much information as possible about the tweet. Therefore, our an-

Sections	Annotation Task
Part 1	**Offensive Language:** None - Weak - Strong
Part 2	**Stance towards the issue.** Pro - Against - Neutral/Not-Applicable
Part 3	**Target Group:** None - Race/Ethnicity - Country/Nationality - Religion - Gender - Sexual Orientation - Opinion Groups
Part 4	**Hate Speech Category:** (Categories indicated in Table 1)

Table 2: The annotation tool has four parts. Annotators were asked about the stance of the tweet about the issue (Istanbul-Convention or Refugees), its offensiveness level; the targeted group of the tweet; and hate speech level.

notation scheme contains four main sections which are illustrated in Table 2.

In the first part, following the work of (Davidson et al., 2017), we asked annotators to separately annotate whether the tweet contains offensive language (all hate speech can be said to be offensive, but not vice versa) or not. A clear definition of offensive speech was not developed, but annotators were just told to annotate a tweet into three categories (*None*, *Weak*, and *Strong*) to indicate the level of offensive language. We asked the annotators to mark offensive language separately, so that a discrimination between offensive language and hate speech can be conducted in the future.

In the second part, annotators were asked about the stance of the tweet about the issue (Istanbul Convention or Refugees). This stance is expressed with 3 options, *Pro*, *Against*, and *Neutral or Not Applicable*.

In the third part, the groups that are targeted in the tweet are annotated if applicable. The following predefined categories are used; Race/Ethnicity, Country/Nationality, Religion, Gender, Sexual Orientation, and Certain Opinion & Status/Position/Position Group.

Finally, the hate speech levels which was set as 0-4, are annotated. The description of the levels and examples are already illustrated in Table 1.

We used LabelStudio[2] which is an open source annotation tool for collecting the annotations. LabelStudio is chosen as it provides different types of annotations performed at the same time.

4. Hate Speech Datasets

We have collected and annotated two hate speech datasets on two different domains using the adopted definition of hate speech and annotation levels described in Section 3.

4.1. Istanbul Convention Dataset

Our data collection contains tweets of the following five days after the withdrawal of Turkey from the Istanbul Convention (between March 20 and 25, 2021). We collected 284,989 tweets in this time period using the top trending topics. The content of the tweets (without considering URLs, hashtags and mentions) were used to remove duplicates. Furthermore, in order to reduce the number of irrelevant tweets, we filtered the examples that contain off-topic hashtags. Among all the remaining tweets collected, 30 popular hashtags about the Istanbul Convention, such as *#istanbulsozlesmesiyasatir* (#IstanbulConventionSavesLives) or *#morardinizmi* (#AreYouPurple/Bruised/Embarassed) were used to identify tweets on the topic of Istanbul Convention. Approximately, 10,000 tweets remained after removing the irrelevant ones.

Following the convention of (Davidson et al., 2017; Poletto et al., 2017), we selected tweets for labeling by including tweets that contain a small number of neutral keywords. These keywords were chosen based on adjectives that contain prejudice against the targeted groups instead of insulting or profanity phrases. We determined them by analyzing the biased discourses of the two sides and applying them equally to both sides (e.g. *Örümcek Kafalı* (Cobweb head), *Kemalist* (Kemalist), *Bağnaz* (Bigot)).

Our 3 annotators are senior undergraduate students in the Cultural Studies department. A total of 1206 tweets in the Istanbul Convention Dataset were labelled by these annotators. Of those tweets, 599 of them were annotated by more than one person.

The Krippendorff alpha agreement scores of the annotators are 0.84 and 0.82 respectively for binary and multi-class settings for 1033 tweets. The distribution of hate speech levels of tweets in the annotated dataset is shown in Table 3.

4.2. Refugee Dataset

In addition to the Istanbul Convention dataset, we collected another dataset on a different domain, in order to analyze the effects of different dataset construction approaches and model performances across domains.

This dataset was collected between January 2020 and September 2021 using Twitter Academic API. As in former studies (Sanguinetti et al., 2018), tweets about immigrants were selected based on commonly used keywords related to immigrants, such as *mülteci* (refugee), *göçmen* (immigrant) and *Suriyeli* (Syrian).

We initially removed the non-Turkish tweets as well as the retweets. In order to eliminate the irrelevant content, we applied the following steps to our tweet collection. Tweets that have more than three hashtags were excluded since accounts using a large number of hashtags might include the trend hashtags to publicize their tweets. Besides, we also kept the tweets which were at least 100 characters long, excluding URLs, hashtags, and mentions. Finally we removed the near duplicate tweets based on their content.

One of the problems with the data we have is the tweets of news about immigrants published by media and journalists' accounts. We have also removed tweets from these sources in order to focus on tweets with emotional/hate speech content. Moreover, tweets from accounts with news-related words such as news, media, agency, and agenda in their usernames were also eliminated. After these steps, we randomly selected a subset of tweets for annotation.

[2]https://labelstud.io

The annotation of this dataset was different than the Istanbul Convention. Hrant Dink Foundation (HDV)[3] was involved in the annotation phase of these tweets besides our three annotators. HDV is working on a project called *Media Watch on Hate Speech* since 2009; in which, all national newspapers and 500 local newspapers are being tracked by HDV's monitoring team; thus they are experts on the topic of hate speech. In the Refugee dataset, HDV members and our annotators annotated 398 and 953 tweets with our annotation scheme, respectively.

An important point is how the choice of keywords used for tweet collection affects the distribution of labels at the end. In Istanbul Convention dataset, we used a more targeted set of keywords and filtering process in order to catch more hate speech, while we performed a random selection within refugee related tweets in order to reduce our bias in the tweet collection process. This step affected the final distribution of these annotations: around 48% tweets are annotated as "Not Hate Speech" in the Istanbul Convention dataset and 60% in the Refugees dataset for non-conflict cases.

Category	# of Tweets	
	IstanbulConv.	Refugee
Not Hate Speech	499	774
Insult	380	181
Exclusion	118	277
Wishing Harm	35	39
Threatening Harm	1	7
Conflicting Annotations	173	-
Tweets w/o Conflicts	1033	1278
Total Tweets	1206	1278

Table 3: Hate speech distribution in the Istanbul Convention Dataset and Refugees Dataset.

4.3. Challenges in Hate Speech Annotation

In addition to the difficulties in coming up with a definition for hate speech, there are other issues in annotating and training models for hate speech, as discussed below. Some of these issues are also highlighted in (Poletto et al., 2017; Sanguinetti et al., 2018).

In the commonly accepted definition, for a tweet to be considered hate speech, it needs to be targeted towards a disadvantaged group. While we hold on to this definition from a social sciences perspective, we believe that dropping the requirement of disadvantaged target group may result in less conflict in the annotations. This is because the target group is sometimes not clear, while the tweet content is hateful towards a general or unknown group, leading to disagreements among annotators.

Hate speech is also *subjective*, due to the complexities of the language and/or cultural perspectives. We have found that some annotators are more inclined to call a tweet as hate speech compared to others. Among the annotators who labelled the same dataset, the minimum and maximum average hate speech level is 0.66 & 0.88 and 0.42 & 0.90 for the Istanbul Convention Dataset and Refugees Dataset, respectively. To address this issue, we have tried to give clear examples in each categories.

Finally, detecting hate speech may require knowing the background; that is something that may not be hate speech on the surface, may be hate speech when one knows the context.

Category	# of Tweets	
	IstanbulConv.	Refugee
None	504	1090
Weak	514	173
Strong	38	82
Missing Value	150	6
Total Tweets	1206	1278

Table 4: Offensive language distribution in the Istanbul Convention Dataset and Refugees Dataset.

Category	# of Tweets	
	IstanbulConv.	Refugee
No Target Group	453	422
Country/Nationality	0	733
Opinion Groups	577	181
Sexual Orientation	168	2
Religion	151	4
Gender	126	9
Race/Ethnicity	27	91
Total Tweets	1206	1278

Table 5: Target group distribution in the Istanbul Convention Dataset and Refugees Dataset. Target group field was allowed to have multiple labels, as there may be multiple targets within one tweet.

Category	# of Tweets	
	IstanbulConv.	Refugee
Pro	745	101
Neutral	46	371
Against	303	863
Missing Value	112	16
Total Tweets	1206	1278

Table 6: Stance distribution in the Istanbul Convention Dataset and Refugees Dataset.

4.4. Target Group, Stance and Offensiveness

In Table 4 and Table 5, the distribution of offensive language and target group labels are shown, respectively. Similar to the hate speech distributions, Istanbul Convention and Refugees Datasets have different distributions on offensive language category and even greater difference can be seen in the target group category. The distribution difference in the target group category was expected, due to domain differences. In Refugees dataset *Country/Nationality* and *Race/Ethnicity* are the groups selected the most. In Istanbul Convention these two are the lowest while the *Sex-*

[3]https://hrantdink.org/en/

Level	Dataset	Approach	Accuracy	F1-Score	Precision	Recall
Binary	Ist. Conv.	Baseline	51.69	51.69	68.16	100.0
		Proposed	**77.06** ($\pm$1.92)	**77.86** ($\pm$2.88)	**77.73** ($\pm$3.05)	**78.33** ($\pm$5.3)
	Refugee	Baseline	39.02	56.14	39.02	100.0
		Proposed	**71.06** ($\pm$4.43)	**65.54** ($\pm$5.27)	**61.66** ($\pm$7.68)	**71.37** ($\pm$8.86)
Multi-Class	Ist. Conv.	Baseline	48.31			
		Proposed	**72.22** ($\pm$2.94)			
	Refugee	Baseline	60.98			
		Proposed	**71.74** ($\pm$2.17)			

Table 7: 5-fold cross validation results for binary and multi-class classification on Istanbul Convention and Refugees Dataset with their average and standard deviations. All metrics are applied with micro averaging. Bold results show improvements over baseline.

ual Orentation, *Religion* and *Gender* are the highest as expected.

In Table 6, the distribution of people's stance on both refugees and the repeal of the Istanbul Convention are given. As with the hate speech distribution, the two datasets have different distributions for stance: while more tweets support the Istanbul Convention, most of the tweets in the Refuge dataset are anti-refugee or neutral.

Since we focused more on the hate speech labels in this paper, we did not exclude samples with missing offensive language, target group or stance labels. We are publicly sharing these annotations as well, together with the dataset.

5. Hate Speech Detection System

In order to create a hate speech detection model, we used two datasets separately for training and evaluated both datasets with their own test sets.

The datasets were annotated in multi-class level. We wanted to try a binary classification setting as well in which no hate speech class becomes the "None" category, and the rest of the other categories are counted as the hate speech class. With this modification we are able to conduct experiments in both binary (hate speech or not) and multi-class (5-class hate speech level) settings.

Furthermore, we trained another model to learn to predict the strength of the hate speech, on a range from [0-4] corresponding to the 5 categories. We believe that the 4 categories of hate speech are roughly on scale, though it is possible that "insult" and "exclusion" categories (categories 1 and 2 respectively) may not be necessarily ordered.

5.1. Models

We fine-tuned a pretrained transformer based model for Turkish, as our baseline model. BERT (Devlin et al., 2018) uses bidirectional transformer architecture for language modeling and represents the state-of-art technique for this problem.

We used the uncased BERTurk model (Schweter, 2020), which was pretrained on a Turkish Wikipedia dump, OSCAR[4], and OPUS[5] data sets, as our BERT encoder which is followed by a sigmoid or softmax layer depending on the task for the classification problem. We used cross-entropy loss function for the classification tasks. For the regression problem, a linear layer is added on top of the encoder, and a mean squared error is applied as a loss function.

5.2. Experimental Setup

We conducted experiments for regression and classification (binary and multi-class) settings, using 5-fold cross validation. The sample size of train, validation and test sets were 70%, 10% and 20%, respectively for each run. Furthermore, we tried 3 different seeds for each fold for model initialization. Hence, there were 15 (3 model seeds and 5 folds) different run for regression, binary and multi-class settings. We also provided results of the majority classifier as the baseline.

Besides, we created test sets for both datasets and reported our results on these sets as a benchmark. Again, models were trained with 3 different seeds.

In the regression problem, since deciding on hate speech is subjective and our classes are on a scale, instead of removing the conflict cases, we averaged the different labels and added them to both datasets. For the classification tasks, we eliminated the conflict cases from the datasets.

5.3. Results

5.3.1. Binary and Multi-Class Classification

Both the binary and multi-class scores for 5-fold cross-validation are reported in Table 7. Since micro average returns the same score for Precision, Recall, Accuracy and F1 in the multi-class case, only one score is reported in the table.

The test results obtained with the Istanbul Convention dataset are very promising, with 77.06% and 72.22% mean accuracy and, 77.86 and 72.22 F1 scores for the binary and multi-class problems, respectively. One of the challenges our model faced was that BERT model was pretrained on formal resources, while our dataset consists of informal and short texts (tweets). Furthermore, Istanbul Convention Dataset is highly imbalanced. Even with these challenges, the obtained results are already very promising (more than 15% and almost 24% over the baseline in accuracy for binary and multi-class respectively).

However, when we trained our model with the Refugees dataset and evaluated in 5-fold cross-validation setting, there was a significant decrease in the performance, even though both models clearly outperformed their respective baselines. Besides the domain difference, there are some other differences between these datasets which may have

[4]https://oscar-corpus.com/

[5]https://opus.nlpl.eu/

Dataset	Approach	RMSE	R^2
Ist. Conv.	Baseline	0.77	-
	Proposed	**0.66** ($\pm$0.05)	0.24 ($\pm$0.09)
Refugee	Baseline	0.96	-
	Proposed	**0.83** ($\pm$0.06)	0.23 ($\pm$0.08)

Table 8: 5-fold cross validation regression results on Istanbul Convention and Refugees datasets with their average and standard deviations.

Dataset	RMSE	R^2
Ist. Conv.	0.67 ($\pm$0.02)	0.21 ($\pm$0.04)
Refugee	0.79 ($\pm$0.04)	0.31 ($\pm$0.01)

Table 9: Official test set regression results on Istanbul Convention and Refugees datasets with their average and standard deviations for 3 different seeds.

caused this decrease. Firstly, the aforementioned differences in dataset collection resulted in a distribution difference. Secondly, the Refugee dataset was annotated by different annotator teams. The difference between background, experience and qualifications of annotators could have led to different annotation behaviors. Both of these are expected behaviors seen across datasets generated in different papers. With these experiments we have replicated a similar setting.

In addition to the cross validation experiments, we created official train-test splits for our datasets. The results of classification with the test portion the two datasets are similar to those obtained with cross-validation and can be seen in Table 10.

5.3.2. Regression Results

The four hate speech classes roughly correspond to a scale from 1 to 4. With this assumption, we applied the same experimental setting and performed regression modelling. The results are shown in Table 8.

Similar to the classification problem, our model performs better in the Istanbul Convention dataset; and both trained models outperform the baselines. However, the results are not very high especially in terms of the R-squared values. We believe that "insult" and "exclusion" categories (categories 1 and 2 respectively) may not be necessarily ordered among each other; while categories 0, 3 and 4 are.

Besides the cross validation scores, we created official train-test splits for our datasets. The results of classification and regression for the two datasets can be seen in Tables 10 and 9.

5.3.3. Error Analysis

The confusion matrix for the multi-class problem is given in Fig. 1 for the official test set of the Istanbul Convention dataset where the accuracy was 76.7%. We see that the main confusion is between adjacent categories (e.g. samples from category 0 labelled as category 1 and vice versa). However, we also observed that our model could not classify any one of the few tweets from categories 3 and 4, which can be attributed to the lack of samples in these categories.

Analyzing individual errors made by our classification and

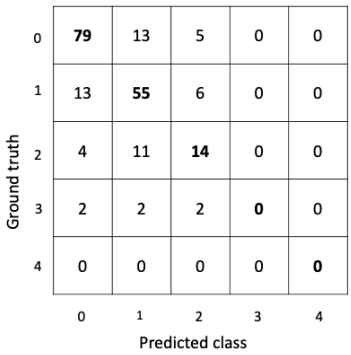

Figure 1: Confusion matrix of the Istanbul Convention dataset in multi-class setting

regression models the Istanbul Convention dataset, we see that the errors are due to several factors: 1) meaning that can be understood from context (i.e. historical or religious references or prior news articles etc); 2) use of contractions (e.g. using only the first letter of a derogatory term); 3) use of infrequent words (e.g. *yallah*); and 4) occasional mistakes in ground-truth.

We give below some sample tweets with corresponding ground-truth and predictions, corresponding to the above issues. In 1), opposition to LGBTQ+ is understood with reference to the people of Lut. In 2), the use of * while writing a derogatory term is not captured by the model, as it is not always used the same way. In 3), the use of infrequent words is not learned by the model. In 4), the tweet is labelled in category 1, but predicted strength is 0.4, which may be considered correct.

1. Turkish: ***lut kavmini helâk etmis*** *rabbimin yolunda anli ak, basi dik yuruyen adam #seninleyizerdogan*

   English: *The man who is walking with honour ...., on the path of lord who* ***destroyed the people of lut*** *#wearewithyouerdogan*

2. Turkish: *mor halkalilar ve* ***l*gbt liler hadi size gule gule*** *#istanbulsozlesmesi*

   English: *Those with purple circles and* ***l*gbtq members, bye bye to you*** *#istanbulconvention*

3. Turkish: *@USER burasi bizim ulkemiz burasi laik turkiye cumhuriyeti bizde burada yasayan vatandaslariz begenmeyen* ***yallah*** *arabistana #istanbulsozlesmesiyasatir #kadincinayetleripolitiktir*

   English: *@USER this is our country, this is the secular republic of turkey, we are citizens living here, if you don't like it,* ***go*** *to arabia #istanbulconventionkeepsalive #femicidesarepolitical*

4. Turkish: *ulkenin cahillik seviyesini gosteren tag #morardinizmi . evet morardik. ama biz sizin kadar kotu degiliz bir gun umarim siz morarmazsiniz.*

Level	Dataset	Accuracy	F1-Score	Precision	Recall
Binary	Ist. Conv.	76.70 ($\pm$1.46)	77.90 ($\pm$1.62)	78.20 ($\pm$1.85)	77.68 ($\pm$3.22)
	Refugee	73.80 ($\pm$0.76)	64.95 ($\pm$1.79)	70.19 ($\pm$1.08)	60.51 ($\pm$3.32)
Multi-Class	Ist. Conv.	71.52 ($\pm$0.28)			
	Refugee	72.34 ($\pm$1.33)			

Table 10: Official test set results for binary and multi-class classification on Istanbul Convention and Refugees datasets with their average and standard deviation for 3 different seeds. All metrics are applied with micro averaging.

English: *#areyoubruised/purple/embarassed hashtag shows the ignorance level of the country. yes we are bruised/purple/embarrassed, but we are not as bad as you, we hope you won't be bruised/purple/embarrassed one day*

6. Conclusions and Future Work

Hate speech definition and annotation are difficult problems. We provide a dataset consisting of two separate subsets on different domains, to foster research on automatic hate speech detection, along with baseline results within these two domains.

Analysis of the errors show that the system is able to capture hate speech that is visible on the surface (e.g. swear words), while missing those that require background knowledge or use of infrequent words or shorthands.

As the first phase of an ongoing project in detecting and measuring hate speech in Turkish, this work highlights the main issues and difficulties in building such a system, from collecting tweets to annotation to detection. We hope that the observations here can guide future work on the topic.

For future work, we have decided to merge categories 3 and 4, so as to have a more balanced dataset. We are also in the process of collecting a much larger dataset in the scope of the ongoing research project, in order to capture the finer and more detailed aspects of the language. We have also observed that hashtags contain a lot of information that would be useful in detecting the stance and also measuring hate speech strength to some level. But since hashtags are domain-specific, the hate speech detection process may benefit from a first pass over the whole collection to understand hashtag polarity.

7. Acknowledgements

This work is supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) (No:119E358).

8. Bibliographical References

ANSA. (2021). Syrians under temporary protection in Turkey top 3.7 million. https://www.infomigrants.net/en/post/35606/syrians-under-temporary-protection-in-turkey-top-37-million. [Online; accessed 06-May-2022].

Bankov, K. (2020). Cyberbullying and hate speech in the debate around the ratification of the istanbul convention in bulgaria: a semiotic analysis of the communication dynamics. *Social Semiotics*, 30(3):344–364.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (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, June. Association for Computational Linguistics.

Burnett, S. (2021). Istanbul Convention: How a European Treaty Against Women's violence became politicized. https://www.dw.com/en/istanbul-convention-how-a-european-treaty-against-womens-violence-became-politicized/a-56953987. [Online; accessed 06-May-2022].

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

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dondurucu, Z. B. and Uluçay, A. P. (2015). The hate speech in new media environments: The analysis of videos which include hate speech for gay people. *International Journal of Social Sciences and Education Research*, 1(3):875–902.

Gagliardone, I., Gal, D., Alves, T., and Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Hatebase.org. (2022). How it Works. https://hatebase.org/how_it_works. [Online; accessed 06-May-2022].

Hüseyin, Y. and Öksüz, O. (2020). Nefret söyleminin inşasında sosyal medyanın rolü: Ekşi sözlük örneği. *Erciyes İletişim Dergisi*, 7(2):1383–1408.

Jacobs, J. B., Potter, K., et al. (1998). *Hate crimes: Criminal law & identity politics*. Oxford University Press on Demand.

Karaman, H. and Işıklı, Ş. (2016). Analysis of religious and ethnic based hate speeches on twitter. *AJIT-e*, 7(25):137.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2020). Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Memisoglu, F. and Ilgit, A. (2017). Syrian refugees in

turkey: multifaceted challenges, diverse players and ambiguous policies. *Mediterranean Politics*, 22(3):317–338.

Poletto, F., Stranisci, M., Sanguinetti, M., Patti, V., and Bosco, C. (2017). Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Sanderson, S. (2021). Turkish people reject Afghan refugees after Taliban takeover. `https://www.infomigrants.net/en/post/34668/turkish-people-reject-afghan-refugees-after-taliban-takeover`. [Online; accessed 06-May-2022].

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Schweter, S. (2020). Berturk - BERT models for Turkish. `https://doi.org/10.5281/zenodo.3770924`. [Online; accessed 06-May-2022].

Walker, S. (1994). *Hate speech: The history of an American controversy*. U of Nebraska Press.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.