JL-Hate: An Annotated Dataset for Joint Learning of Hate Speech and Target Detection

Kaan Buyukdemirci[1], **Izzet Emre Kucukkaya**[2]*, **Eren Olmez**[1], **Cagri Toraman**[3]*
[1]Department of Electrical and Electronical Engineering, Bilkent University, Ankara, Turkey
[2]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[3]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey
{kaan.buyukdemirci, eren.olmez}@ug.bilkent.edu.tr
emre.kuecuekkaya@tum.de
ctoraman@ceng.metu.edu.tr

Abstract

The detection of hate speech is a subject extensively explored by researchers, and machine learning algorithms play a crucial role in this domain. The existing resources mostly focus on text sequence classification for the task of hate speech detection. However, the target of hateful content is another dimension that has not been studied in details due to the lack of data resources. In this study, we address this gap by introducing a novel tweet dataset for the task of joint learning of hate speech detection and target detection, called JL-Hate, for the tasks of sequential text classification and token classification, respectively. The JL-Hate dataset consists of 1,530 tweets divided equally in English and Turkish languages. Leveraging this dataset, we conduct a series of benchmark experiments. We utilize a joint learning model to concurrently perform sequence and token classification tasks on our data. Our experimental results demonstrate consistent performance with the prevalent studies, both in sequence and token classification tasks.

Keywords: Hate speech detection, Joint learning, Target detection

Bias Statement: *This paper discusses examples of harmful content and hate speech stereotypes. The authors do not support the use of harmful language, nor any of the harmful representations quoted below.*

1. Introduction

Online hate speech has detrimental effects on individuals and societies (Kelly et al., 2018). It may serve as an early indicator of more severe offenses, given that numerous instances of hate attacks have occurred subsequent to the online expression of hateful content by the perpetrator or suspect (Robertson et al., 2018; Times, 2019). Hate speech datasets play an important role in advancing research and technology aimed at the detection and comprehension of hate speech, potentially contributing to the prevention of such offenses.

The datasets serve as the foundation upon which machine learning models are trained, validated, and evaluated to develop effective hate speech detection systems (Yadav et al., 2023). The availability of diverse and representative hate speech datasets is crucial for creating models that can generalize across various linguistic, cultural, and contextual dimensions. By publishing datasets that encompass a wide range of hate speech instances, researchers can uncover underlying patterns, linguistic markers, and contextual cues that aid in the identification of hate speech content.

Despite the growing awareness of the challenges posed by hate speech, there remains a significant gap in the availability of hate speech datasets (MacAvaney et al., 2019), particularly in two aspects. First, existing resources mostly focus on text sequence classification (Poletto et al., 2021). However, the target of hateful content is as important as detecting hateful text sequences, since the analysis of hateful content could differ with respect to different target groups. Second, existing resources mostly support English. This scarcity of resources poses a barrier to effectively addressing hate speech in linguistic contexts that differ from the dominant English-language datasets. One such case is the Turkish-language datasets. Turkish content, as a linguistically and culturally rich context, presents unique challenges in hate speech detection that cannot be adequately addressed without dedicated Turkish datasets (Toraman et al., 2022). The absence of comprehensive hate speech datasets including hate targets hinders the development of accurate and contextually relevant hate speech detection models for this language.

The primary motivation of this study is to contribute to the global effort by introducing a novel hate speech dataset for the task of joint learning of hate speech and target detection, called **JL-Hate** (Joint Learning Hate Speech Dataset)[1], comprising

*Work partially done in Aselsan, Ankara, Turkey.

[1]We publish the JL-Hate dataset and source codes for the benchmark experiments at https://github.com/metunlp/JL-Hate

LREC-COLING 2024, pages 9543–9553
20-25 May, 2024.

Neutral Speech	I hate Balloon x Baseball Who tf is outside making all this damn noise?? Are they playing basketball?????
Offensive Speech	all arabs do is get mad at something stupid then become rude with everyone around them Fuck white privilege. There, I said it. Fuck you white people Fuck white beauty standard Fuck you Trump
Hate Speech	Every single time a gay says there gay I'm going to Tie a noose and kill them "I hate being a part of the generation that can't figure out what their gender is" okay, then die

Table 1: Examples of various types of speech on social media.

1,530 tweets, divided equally to English and Turkish and specifically tailored for sequence and token classification tasks.

Recent regulatory changes have made the acquisition of tweets more challenging for research and dataset construction (Calma, 2023). Given these challenges and constraints, this dataset gains added significance as a valuable resource for addressing this critical issue. JL-Hate empowers researchers and practitioners to devise more effective hate speech detection tools that account for the distinctive linguistic and cultural characteristics of the Turkish language. Furthermore, we extend JL-Hate and introduce an English Hate Speech dataset, comprising once again 765 tweets annotated by the same annotators, aiming to engage a broader audience and facilitate comparisons with other academic research.

The main contributions of this study are that we **(i)** publish a novel dataset for joint learning of hate speech and target detection in English and Turkish, **(ii)** provide a brief summary of related work on joint learning for hate speech detection, and **(iii)** conduct benchmark experiments on this novel dataset with a detailed error analysis and a comparison with other related studies.

2. Background

Definition of Hate Speech in This Study According to the definition provided by United Nations (2023), hate speech encompasses "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor." However, it should be noted that a universally accepted definition of hate speech is yet to be established under international human rights law. The concept remains subject to ongoing discussions, particularly concerning freedom of opinion and expression, non-discrimination, and equality.

In light of this definition, and inspired by Toraman et al. (2022), we establish the following definitions: Hate speech is characterized by posts that specifically target, incite violence against, threaten, or call for physical damage for an individual or a group of people because of some identifying trait or characteristic. Offensive speech is characterized by posts that humiliate, taunt, discriminate, or insult an individual or a group of people in any form. Neutral speech pertains to posts that do not fall into either of the preceding categories. Illustrative examples of posts conforming to these definitions can be found in Table 1.

Sequence Classification Sequence classification is a pivotal technique in natural language processing, enabling the categorization of entire text sequences into predefined classes. Within our research framework, sequence classification models are used to categorize each text as hateful, offensive, or neutral (Table 1).

Token Classification Token classification, which mainly focuses on span detection and named entity recognition, is centered around pinpointing distinct elements within a text, like individual words or phrases, that align with specific categories or entities. Within our research framework, token classification models serve to identify the signal – the segment responsible for rendering the text hateful or offensive – as well as the target – the recipient of the hateful or offensive speech (Table 6).

Joint Learning Joint learning entails training a single model to simultaneously perform multiple interconnected tasks. In our study, this approach is applied to address sequence classification and token classification concurrently. By capitalizing on shared knowledge and representations across these related tasks, Joint Learning enhances the model's overall comprehension of hate speech. This integrated strategy can contribute to a more accurate analysis of hate speech content (Jeong et al., 2022).

3. Related Work

3.1. Datasets for Hate Speech Detection

The field of hate speech research extensively explores sequence classification. Notable studies include the work of Davidson et al. (2017), who introduce a new dataset of 24,802 Twitter posts in English, employing Support Vector Machine (SVM)

Study	Size	Lng	Domain	Seq			Tok		
				N	O	H	T	S	JL
Davidson et al. (2017)	24,802	En	Twitter	✓	✓	✓			
Zampieri et al. (2019)	14,100	En	Twitter	✓		✓			
Luu et al. (2021)	33,400	Vi	Facebook, YouTube	✓	✓	✓			
Zhu et al. (2021)	10,617	En	News					✓	
Mathew et al. (2021)	20,148	En	Twitter, Gab	✓	✓	✓		✓	✓
Beyhan et al. (2022)	2,484	Tr	Twitter	✓	✓	✓			
Toraman et al. (2022)	200k	En, Tr	Twitter	✓	✓	✓			
Zhou et al. (2022)	10,800	En	Twitter, hate forums	✓		✓		✓	
Pavlopoulos et al. (2022)	11,006	En	News					✓	
Jeong et al. (2022)	40,429	Ko	News, YouTube	✓	✓		✓	✓	✓
Markov and Daelemans (2022)	6,000	Nl	Facebook						
Hoang et al. (2023)	11,056	Vi	Facebook, YouTube					✓	
This Study	1,530	En, Tr	Twitter	✓	✓	✓	✓	✓	✓

Table 2: **A Brief Summary of the Existing Studies for Hate Speech Detection in the Literature** The following is a compilation of studies, wherein the abbreviations 'N', 'O', and 'H' represent 'Neutral', 'Offensive', and 'Hateful', respectively, denoting the content categories of the dataset utilized for sequence classification. Similarly, the abbreviations 'T' and 'S' represent 'Target' and 'Signal', respectively, indicating the content categories of the dataset employed for token classification. Lastly, the abbreviation 'JL' signifies 'Joint Learning', indicating whether any joint learning methodology has been employed.

for classifying each tweet as neutral, offensive, or hateful. Another study by Zampieri et al. (2019) introduces the OLID dataset containing 14,100 English Twitter posts, classified in terms of offensiveness, targetedness, and the target's categorization. They utilize a Convolutional Neural Network (CNN) model based on the architecture proposed by Kim (2014). Additionally, Luu et al. (2021) introduce the ViHSD dataset consisting of 33,400 Vietnamese comments. They employ the Multilingual-Bert-Base (Devlin et al., 2019) model to classify comments as clean (neutral), offensive, or hateful. Furthermore, Beyhan et al. (2022) present a Turkish dataset where Twitter posts are categorized in four different aspects, including offensive language, stance towards the issue, target group, and hate speech category. They achieve their results using the BERTurk (Schweter, 2020) architecture. Lastly, Toraman et al. (2022) introduce a new dataset consisting of 200,000 tweets in English and Turkish, employing state-of-the-art methods such as Megatron (Shoeybi et al., 2019) for English and ConvBERTurk (Schweter, 2020) for Turkish. They classify each text as neutral, offensive, or hateful.

3.2. Datasets for Target Detection

Token classification is another prominent research area within the hate speech domain. Zhu et al. (2021) employ an ensemble method to classify tokens as empty or signal, achieving the first rank in the SemEval-2021 Task 5 competition (Pavlopoulos et al., 2021) on the CivilComments (Borkan et al., 2019) dataset. Many other studies are involved in SemEval-2021 (Kotyushev et al., 2021; Gia Hoang et al., 2021; Hossain et al., 2021; Khan et al., 2021; Salemi et al., 2021; Luu and Nguyen, 2021) as summarized by Ravikiran et al. (2022). Additionally, Zhou et al. (2022) combines data across multiple platforms to the size of 10,800. They then employ Support Vector Classifier (Boateng et al., 2020) and BERT (Devlin et al., 2019) to achieve their best results. Also, Pavlopoulos et al. (2022) introduces ToxicSpans, a new dataset of 11,006 comments from CivilComments dataset. They employ SPAN-BERT-SEQ (Joshi et al., 2020) to achieve their best result. Furthermore, Hoang et al. (2023) extend the work of Luu et al. (2021) and introduce the ViHOS dataset, incorporating tokens labeled as empty or signal. They achieve their best results using XLM-RoBERTa (Conneau et al., 2020) and PhoBERT (Nguyen and Tuan Nguyen, 2020). Another study by Markov and Daelemans (2022) utilize 6,000 Dutch Facebook comments from Markov et al. (2021) to classify each hate speech target as migrants or other.

3.3. Joint Learning

Some studies adopt joint learning approaches to simultaneously address both sequence classification and token classification tasks. Mathew et al. (2021) introduce the HateXplain dataset, containing 20,148 English Twitter posts. They use a weighted joint loss function to jointly learn for both tasks at the same time. Another study by Jeong et al. (2022) introduce the KOLD dataset comprising 40,429 Korean comments, utilizing RoBERTa (Liu et al., 2019) to jointly handle both sequence and token classification. The area of joint learning in the context of

Definition	EN	TR
Number of Tweets	765	765
Number of Neutral Tweets	334	258
Number of Offensive Tweets	277	349
Number of Hateful Tweets	54	67
Number of Tweets with Hashtags	155	292
Number of Tweets with URLs	241	210
Number of Tweets with Emojis	74	64
First Tweet Year	2020	2020
Last Tweet Year	2021	2021
Shortest Tweet Length in Words	5	5
Longest Tweet Length in Words	59	48
Number of Users	761	758
Labeled by two annotators	581	586
Labeled by four annotators	184	179

Table 3: Main statistics of the JL-Hate dataset.

hate speech is more limited compared to sequence and token classification.

3.4. Languages

While the majority of existing studies in this field predominantly focus on the English language (Davidson et al., 2017; Zampieri et al., 2019; Zhu et al., 2021; Kotyushev et al., 2021; Gia Hoang et al., 2021; Hossain et al., 2021; Khan et al., 2021; Salemi et al., 2021; Luu and Nguyen, 2021; Mathew et al., 2021; Toraman et al., 2022; Zhou et al., 2022; Pavlopoulos et al., 2022), several studies have also explored hate speech in various other languages, such as Vietnamese (Luu et al., 2021; Hoang et al., 2023), Korean (Jeong et al., 2022), and Turkish (Toraman et al., 2022; Beyhan et al., 2022).

3.5. Our Differences

For a comprehensive overview of the discussed studies, please refer to Table 2. Moreover, there are more in-depth reviews of the existing literature available (Poletto et al., 2021). To the best of our knowledge, no existing study has specifically focused on token classification in the Turkish language. The primary aim of this study is to fill this gap in the existing literature by presenting a new dataset. Furthermore, we enhance the current body of research by providing a dataset that involves target token classification in both English and Turkish.

4. Dataset

4.1. Dataset Contents and Statistics

We provide an overview of our dataset, JL-Hate, derived from Toraman et al. (2022). JL-Hate comprises 765 English and 765 Turkish tweets. These tweets are evenly distributed across various topics such as religion, gender, race, politics, and sports. Moreover, the dataset ensures a balanced representation of labels, covering neutral, offensive, and hateful content. We also consider the average confidence scores of annotations in terms of 0.6, 0.8, and 1.0 (the higher the more confident), as detailed in Toraman et al. (2022). To achieve this, we randomly sample 17 tweets from each of five topics (i.e. religion, gender, race, politics, and sports), with three classes (i.e. neutral, offensive, and hate), three confidence levels (i.e. 0.6, 0.8, 1.0), and two languages (i.e. English and Turkish), totaling 1,530 tweets.

We also discard the sequence-level annotations provided by Toraman et al. (2022), and ask our annotators to label both sequence and span-level, ensuring that our annotators are not influenced by previous annotations.

Each individual tweet in the dataset undergoes a rigorous annotation process, involving two annotators who evaluate the content both at the text-level and the span-level. In cases where discrepancies arise at the text-level annotation, an additional round of assessment is conducted by two more annotators, again covering both text-level and span-level evaluations.

Table 3 provides a comprehensive summary of the dataset's key characteristics, encompassing various aspects such as tweet statistics, hashtag usage, URL inclusion, emoji presence, date range, tweet length, user count, and inter-annotator agreements. Three of the annotators are male graduate students in mid-20s, and the remaining annotator is a male senior researcher in late-30s. The average Cohen's Kappa is calculated to be 0.418 and it represents the average value across each pairs.

4.2. Text-Level Annotations

In this section, we delve into the process of text-level annotations. Table 4 provides a comprehensive breakdown of the dataset across different domains, offering insights into the distribution of annotations.

Each tweet within our dataset undergoes initial evaluation by two annotators who classify it into one of three categories: neutral, offensive, or hateful. In instances where discrepancies arise between the annotators, a secondary round of annotation is initiated, again involving two additional annotators. Consequently, each tweet is assigned one of five classification labels: "Skipped," "Tie," "Neutral," "Offensive," or "Hateful."

A tweet is categorized as "Skipped" if an annotator intentionally refrains from providing an annotation, a mechanism designed to exclude ambiguous or politically charged tweets from our dataset. Alternatively, a tweet may fall into the "Tie" category

Lang.	Domain	S	T	N	O	H	Total
EN	Religion	18	3	72	49	11	153
	Gender	27	6	63	38	19	153
	Race	15	5	56	66	11	153
	Politics	7	6	59	74	7	153
	Sports	9	4	84	50	6	153
TR	Religion	21	5	44	66	17	153
	Gender	16	3	53	65	16	153
	Race	17	5	52	68	11	153
	Politics	12	2	51	76	12	153
	Sports	2	8	58	74	11	153

Table 4: **Distribution of Dataset Across Content Domains and Annotation Categories** The abbreviations 'S', 'T', 'N', 'O', and 'H' represent 'Skipped', 'Tie', 'Neutral', 'Offensive', and 'Hateful', respectively, denoting the content categories of the dataset.

	EN		TR	
Definition	**O**	**H**	**O**	**H**
Number of Tweets with HTARs	0	45	0	53
Number of Tweets with HSIGs	0	50	0	66
Number of Tweets with OTARs	230	5	280	5
Number of Tweets with OSIGs	259	13	341	30
Avg. Length of HTAR in Words	0	3	0	3
Avg. Length of HSIG in Words	0	3	0	3
Avg. Length of OTAR in Words	2	3	2	3
Avg. Length of OSIG in Words	2	3	2	2

Table 5: **Summary of Span-Level Annotations Across Content Domains and Languages** The abbreviations 'Avg', 'Ct', 'HTAR', 'HSIG', 'OTAR', 'OSIG', 'EN', and 'TR' represent 'Average', 'Count', 'Hate Target', 'Hate Signal', 'Offense Target', 'Offense Signal', 'English', and 'Turkish' respectively.

if there is no majority agreement among the four annotations.

4.3. Span-Level Annotations

This section provides an in-depth look at our span-level annotation process, an essential component of our dataset preparation. Table 5 gives a comprehensive summary of span-level annotations across both English (EN) and Turkish (TR) languages.

Following the completion of text-level annotations, we proceed to merge span-level annotations based on the majority text-level annotation. Specifically, if an annotator's text-level annotation differs from the majority, their span-level annotations are not considered in the merging process. This approach ensures the consistency of span-level labels within each category.

To facilitate the combination of diverse span-based annotations, we assign priority levels to each token category, with the hierarchy as follows: HTAR (Hate Target) > HSIG (Hate Signal) > OTAR (Offense Target) > OSIG (Offense Signal). Notably, we prioritize H (Hateful) over O (Offensive) tokens in response to the dataset's inherent imbalance, as demonstrated in Table 5.

Algorithm 1 Merge Span-Level Annotations

Given: $mtla$ (Majority Text-Level Annotation), tla_i (Text-Level Annotation), and sla_i (Span-Level Annotation) for $i \in [1, 4]$
Initialize $combined\_span\_annotations$ as an empty list
if $mtla = S$ **or** $mtla = T$ **then**
  return $combined\_span\_annotations$
end if
for $priority$ **from** high priority **to** low priority **do**
  Retrieve spans with priority level $priority$ from sla_i where $tla_i = mtla$ for $i \in [1, 4]$
  Discard spans that intersect with any span in $combined\_span\_annotations$
  Apply union operation to the remaining spans (e.g., "stupid gay" union "gay people" = "stupid gay people")
  Add the remaining spans to $combined\_span\_annotations$
end for
return $combined\_span\_annotations$

Furthermore, we accord priority to TAR (Targeted) over SIG (Signature) tokens to address the imbalance between these two tokens, as demonstrated in Table 5. Adhering to these priority levels, we execute the annotation combination process through the algorithm outlined in Algorithm 1. This algorithm carefully merges span-level annotations, taking into account the established priorities, resulting in a consistent and coherent dataset for our subsequent analyses.

4.4. Tokenization and Tagging

Following the merge, we engage in a tokenization process for the tweets. Each token's label is determined by associating it with the highest-priority span within its designated token range. Table 6 provides some examples resulting from this process.

Moreover, we implement the IO (Inside, Outside) tagging scheme in our dataset's structure. While the more prevalent tagging scheme is BIO (Beginning, Inside, Outside), due to our limited data availability and the added complexity of introducing more labels with BIO tagging, we opt for IO tagging. The decision is further motivated by the fact that our data is not tokenized to accommodate the BIO tagging convention.

Every single time a gay says there gay I'm going to Tie a noose and kill them
"I hate being a part of the generation that can't figure out what their gender is" okay, then die
all arabs do is get mad at something stupid then become rude with everyone around them
Fuck white privilege . There, I said it. Fuck you white people Fuck white beauty standard Fuck you Trump
Hate Target Hate Signal Offense Target Offense Signal

Table 6: Illustrative examples for token classification (target detection).

5. Experiments

5.1. Experimental Setup

In our experimental setup, we employ a 0.9 split ratio to allocate data for the training set, reserving 0.1 of the dataset for evaluation. To ensure robustness and a comprehensive assessment of our model, we adopt a 10-fold cross-validation approach. This entails training our model 10 times, each time with a distinct split for training and evaluation, iterating through varying subsets of the data.

The best performing model, as defined by this criterion, is chosen from each of 10 runs. Subsequently, we aggregate the results from these runs to calculate the mean and standard deviation for each performance metric. We report Macro F1 score for both sequence and token classification. We also report evaluation scores for individual classes.

For our experiments on the Turkish language, we utilize the ConvBERT model[2] (Jiang et al., 2020) with 12 layers, 768 dimension and 12 heads, a total of 106M parameters. For English, we opt for the DistilRoBERTa model[3] (Liu et al., 2019; Sanh et al., 2019) with 6 layers, 768 dimension and 12 heads, a total of 82M parameters. Following thorough experimentation with various models, we ascertain that these two models exhibit commendable performance and also demonstrate memory-efficiency, as demonstrated by the works of Toraman et al. (2023); Liu et al. (2019); Sanh et al. (2019). It's also noteworthy to mention that we utilize the HuggingFace[4] and PyTorch[5] framework to facilitate our experimentation process. Additionally, we open-source our implementation, making it publicly accessible[6].

We use AdamW optimizer with default hyperparameters for both models. Learning rate is 0.001 and weight decay is 0.01. Input sequence length is set to 128. We conduct training for 30 epochs and select the best performing model based on the highest token classification Macro F1 score.

As illustrated in Figure 1, we employ joint learning to execute sequence and token classification tasks jointly. Upon the language models utilized in this study (i.e. ConvBERT and DistilRoBERTa), we have two separate single feed-forward layers for sequence and token classification, respectively. The final loss function is based on a weighted sum of sequence classification loss (cross-entropy) and token classification loss (cross-entropy), with weights of 0.1 and 0.9, respectively. We empirically select the weights based on our preliminary experiments, where we observe better performance. The higher weight on token classification can be attributed to the dataset's abundance of empty tokens, which biases models towards predicting mostly the empty label.

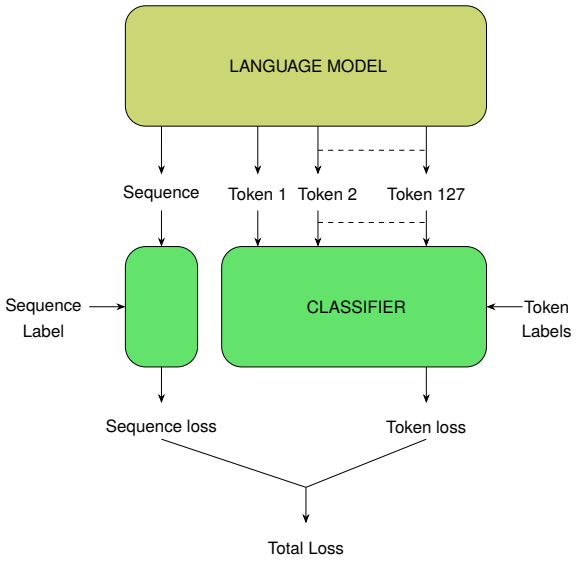

Figure 1: **An illustration for joint learning of hate speech and target detection.** The yellow block represents the encoder-based language model, and the green blocks represent the classifiers (feed-forward layers).

5.2. Benchmark Results

The detailed results of our study are presented in Table 7. To establish a comparative framework for our results, we summarize the findings of the studies discussed in the related work section (refer to Table 8). It is important to note that each study varies in terms of experiment setups, models,

[2] https://huggingface.co/dbmdz/convbert-base-turkish-cased
[3] https://huggingface.co/distilroberta-base
[4] https://github.com/huggingface
[5] https://github.com/pytorch/pytorch
[6] https://github.com/metunlp/JL-Hate

	Sequence F1 (%)				Token F1 (%)					
	Macro F1	Neutral F1	Offensive F1	Hateful F1	**Macro F1**	HTAR F1	HSIG F1	OTAR F1	OSIG F1	O F1
English	**68.5** ± **7.8**	77.4 ± 7.3	77.0 ± 6.9	51.2 ± 18.2	**49.6** ± **8.6**	21.0 ± 26.0	29.8 ± 18.2	41.5 ± 9.0	59.1 ± 6.2	96.4 ± 0.6
Turkish	**71.7** ± **5.3**	74.8 ± 5.3	82.1 ± 5.8	58.1 ± 16.9	**52.3** ± **5.3**	32.5 ± 14.6	34.5 ± 14.7	40.8 ± 5.6	59.7 ± 3.8	94.0 ± 0.8

Table 7: **Benchmark Experimental Results** The following are the averages and standard deviations of results obtained by 10-fold cross-validation, wherein the abbreviations 'HTAR', 'HSIG', 'OTAR', 'OSIG', and 'O' represent 'Hate Target', 'Hate Signal', 'Offense Target', 'Offense Signal', and 'Empty', respectively, denoting the different token classes utilized for token classification.

Studies	Models	Metric	Seq (%)	Tok (%)	Joint (%)
Davidson et al. (2017)	SVM	Overall F1	51	-	-
Zampieri et al. (2019)	CNN	Macro F1	47	-	-
Luu et al. (2021)	BERT	Macro F1	63	-	-
Zhu et al. (2021)	Ensemble	F1	-	71	-
Mathew et al. (2021)	BERT-HateXplain	Macro F1	-		69
Beyhan et al. (2022)	BERTurk	Micro F1	78, 66	-	-
Toraman et al. (2022)	Megatron, ConvBERTurk	F1	82, 78	-	-
Zhou et al. (2022)	SVC, BERT	F1	67, 41, 59	68	-
Pavlopoulos et al. (2022)	SPAN-BASED-SEQ	F1	-	63	-
Jeong et al. (2022)	RoBERTa	F1	77, 58	52, 72	-
Markov and Daelemans (2022)	BERTje (De Vries et al., 2019)	F1	69	-	-
Hoang et al. (2023)	XLM-RoBERTa, PhoBERT	F1	-	78, 69	-

Table 8: **The Performance Results from Other Studies** 'Seq', 'Tok', and 'Joint' refers to the sequence, token, and joint performances for the given model and metrics, respectively.

metrics, and the nature of the tasks, making direct ranking impractical. Therefore, the purpose of this table is to offer a broad comparative perspective rather than rank the studies.

5.3. Sequence Classification

Referring to Table 8, despite the modest size of our dataset, our models perform within a comparable range to other studies concerning sequence classification performance. For a direct comparison aligned with the same sequence classification setup, metrics, and task, we can specifically consider the work by Toraman et al. (2022). Both studies involve classifying tweets into neutral, offensive, or hateful categories based on a shared definition and dataset.

The study by Toraman et al. (2022) achieves 82% macro F1 for English and 78% macro F1 for Turkish tweets. Notably, this discrepancy in performance can be attributed primarily to the dataset size, as they utilized approximately 130 times more tweets.

A closer examination of Table 7 reveals that we attain higher F1 scores for neutral and offensive tweets, while registering a lower F1 score for hateful tweets. This observation is consistent with the imbalances evident in the class distribution outlined in Table 3. Hateful tweets constitute only about 8% of English and 10% of Turkish tweets, whereas offensive and neutral tweets are distributed almost evenly among the remaining tweets.

5.4. Token Classification

Drawing from Table 8, our models demonstrate performance within a comparable range to other studies in token classification, despite the constrained size of our dataset. A study closely related to our work in terms of token classes is that of Jeong et al. (2022). They achieve an F1 score of 52% for offense signal and 72% for offense target tokens.

It is essential to emphasize that, as additional token labels, we differentiate between hate signal and hate target tokens, which impacts our macro F1 score as well as the F1 score for each token class. An intriguing observation is that Jeong et al. (2022) achieved superior target span detection compared to signal detection, while our results show the opposite trend. This discrepancy might stem from the distinct linguistic characteristics of their Korean dataset compared to our dataset.

A deeper understanding of why we excel in signal detection over target detection and in detecting hate better than offense can be understood from Table 5. The class imbalance is a significant factor influencing this pattern, as there are notably more signals than targets, and offense instances outweigh hate instances by a factor of approximately 5 to 1.

6. Error Analysis

Understanding errors is essential in our study, as it provides valuable insights into the factors contributing to the suboptimal performance of our hate speech detection model. By analyzing these errors,

Text	Sequence Label	Prediction
No, I'm not bi. No, I'm not gay/ lesbian. No, I'm not pan. No, I'm not straight. Fucking leave me alone, dude, I don't use labels and that's the fucking shit. You're still valid even if you don't decide to label your sexuality	Neutral Speech	Offensive Speech
Please justify why George Floyd had to die by having a knee on his neck. Why the fuck is there even a question on this? What kind of society have we become? Hey, Church... WHERE ARE WE?	Neutral Speech	Offensive Speech
Miss me with all that church shit. I'm good.	Offensive Speech	Neutral Speech
#HamidAnsari was not able to be dept. President of india. He was able to be peon of my school. Terrorist supporter and love. Throw him out from india. ...	Hate Speech	Offensive Speech

Table 9: **Examples for Sequence Classification Errors**

Token Labels	Name a better duo than Manchester United and nearly giving fans a heart attack! ...
Token Predictions	Name a better duo than Manchester United and nearly giving fans a heart attack! ...
Token Labels	Why is #israel allowed to play international football? Is not Kick it Out! Against racism? Will Football players not take the knee for the thousands of indigenous #Palestinians killed. Ethnically cleansed, #FIFA you are a Hypocrite.
Token Predictions	Why is #israel allowed to play international football? Is not Kick it Out! Against racism? Will Football players not take the knee for the thousands of indigenous #Palestinians killed. Ethnically cleansed, # FIFA you are a Hypocrite.

Table 10: **Examples for Token Classification Errors** The same coloring scheme is used as in Table 6.

we can devise strategies for enhancing the model's performance. The instances of sequence errors are accessible in Table 9, while instances of token errors are documented in Table 10.

6.1. Sequence Errors

In the first example, the text expresses the author's personal opinion without disseminating offensive or hateful content or targeting any specific group. Despite this, the model inaccurately classifies it as offensive speech. This can be attributed to the subject matter's sensitivity and the inclusion of profane language by the author.

Similarly, in the second example, the author employs strong language to voice their disapproval of a situation, but their expression does not involve the endorsement of offensive or hateful content or any discriminatory intent towards a particular group. Nevertheless, the model inaccurately classifies as offensive speech, which is likely influenced by the topic's sensitivity.

The third example involves the author discussing a group's activities in a disrespectful manner. The model's misclassification of this example as neutral speech can be attributed to the limited amount of training data that addresses the nuances of this form of offensive speech, as typical offensive speech tends to be more straightforward in nature.

In the fourth instance, the author appears to endorse a forceful action directed at an individual, a categorization that aligns with our definition of hate speech. Misclassification in this case may result from its tendency to classify samples as offensive rather than as hate speech, stemming from the sample class imbalance, as illustrated in Table 3.

6.2. Token Errors

In the first example, the model's misclassification of the term "Manchester United" as an offensive target may appear mysterious initially. However, it is plausible that the model inaccurately interpret the author's emotional tone as anger towards "Manchester United" due to the expression "giving fans a heart attack!" while overlooking the underlying sarcasm.

The second example demonstrates the complexities involved in annotating certain tweets, as opinions on whether the author is merely expressing their viewpoint or engaging in offensive behavior may differ among readers. In this case, the model leans towards the latter interpretation, as evidenced by its classification of "FIFA" as offense target and "Hypocrite" as offense signal.

7. Conclusion

We introduce a joint learning dataset, called JL-Hate, comprising a total of 1,530 tweets, evenly divided to English and Turkish. The dataset focuses on text sequence and token classification tasks for hate speech. In order to form this comprehensive dataset, we followed a detailed annotation process utilizing both text-level and span-level annotations. After merging span-level annotations,

we performed tokenization based on the priority of each span. During this tokenization process, we use the IO tagging scheme.

In our efforts to establish a benchmark understanding, we provide the baseline performance results for our dataset. Furthermore, we present a detailed error analysis on the prediction results of sequence and token classification, deepening the insights into the experiments.

In future work, we plan to extend the dataset to more instances and languages for diversity. Another dimension can be the relation extraction or detection between different entities in hateful content, such as sources and targets.

8. Limitations and Ethical Concerns

Creating a hate speech dataset can be more difficult due to the regulations of social media platforms. Making the dataset balanced in terms of labels can be therefore challenging, though we give much effort to make it as balanced as possible.

Human annotation is a costly and laborious process for specifically token classification (target detection). The annotators were given careful guidelines on the definitions of class labels. However, the dataset can still reflect their personal biases and interpretations to some extent.

We acknowledge the relatively smaller size of our dataset. However, our experiments demonstrate that despite its size, our dataset still yields results comparable to those reported in related studies within the field. Our study focuses on the English and Turkish languages only, which might reflect the cultural biases.

Another critical aspect to emphasize is the core intention of this study. Our primary objective is not to label individuals, but rather to enhance our understanding of hate speech. This deeper comprehension serves as a foundational step toward implementing preventive measures.

9. Bibliographical References

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.

Ernest Yeboah Boateng, Joseph Otoo, and Daniel Abaye. 2020. Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, 08:341–357.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Justine Calma. 2023. Twitter just closed the book on academic research. [Accessed on 24-Mar-2024].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 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(1), pages 512–515.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

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.

Phu Gia Hoang, Luan Thanh Nguyen, and Kiet Nguyen. 2021. UIT-e10dot3 at SemEval-2021 task 5: Toxic spans detection with named entity recognition and question-answering approaches. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 919–926, Online. Association for Computational Linguistics.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–

669, Dubrovnik, Croatia. Association for Computational Linguistics.

Tashin Hossain, Jannatun Naim, Fareen Tasneem, Radiathun Tasnia, and Abu Nowshed Chy. 2021. CSECU-DSG at SemEval-2021 task 5: Leveraging ensemble of sequence tagging models for toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 990–994, Online. Association for Computational Linguistics.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33:12837–12848.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yvonne Kelly, Afshin Zilanawala, Cara Booker, and Amanda Sacker. 2018. Social media use and adolescent mental health: Findings from the UK millennium cohort study. *EClinicalMedicine*, 6:59–68.

Yakoob Khan, Weicheng Ma, and Soroush Vosoughi. 2021. Lone pine at SemEval-2021 task 5: Fine-grained detection of hate speech using BERToxic. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 967–973, Online. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Mikhail Kotyushev, Anna Glazkova, and Dmitry Morozov. 2021. MIPT-NSU-UTMN at SemEval-2021 task 5: Ensembling learning with pre-trained language models for toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 913–918, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 415–426. Springer.

Son T. Luu and Ngan Nguyen. 2021. UIT-ISE-NLP at SemEval-2021 task 5: Toxic spans detection with BiLSTM-CRF and ToxicBERT comment classification. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 846–851, Online. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Ilia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35(17), pages 14867–14875.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

1037–1042, Online. Association for Computational Linguistics.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:1–47.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha S, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on offensive span identification fromCode-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 261–270, Dublin, Ireland. Association for Computational Linguistics.

Campbell Robertson, Christopher Mele, and Sabrina Tavernise. 2018. 11 killed in synagogue massacre; suspect charged with 29 counts.

Alireza Salemi, Nazanin Sabri, Emad Kebriaei, Behnam Bahrak, and Azadeh Shakery. 2021. UTNLP at SemEval-2021 task 5: A comparative analysis of toxic span detection using attention-based, named entity recognition, and ensemble models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 995–1002, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Stefan Schweter. 2020. Berturk - bert models for turkish. *GitHub*. [Accessed on 24-Mar-2024].

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

The New York Times. 2019. Christchurch shooting live updates: 49 are dead after 2 mosques are hit. [Accessed on 24-Mar-2024].

Cagri Toraman, Izzet Emre Kucukkaya, Oguzhan Ozcelik, and Umitcan Sahin. 2023. Tweets under the rubble: Detection of messages calling for help in earthquake disaster. *arXiv preprint arXiv:2302.13403*.

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.

United Nations. 2023. What is hate speech? https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech. [Accessed on 24-Mar-2024].

Arun Kumar Yadav, Mohit Kumar, Abhishek Kumar, Shivani, Kusum, and Divakar Yadav. 2023. Hate speech recognition in multilingual text: hinglish documents. *International Journal of Information Technology*, 15.

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.

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2022. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, page 1–28.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.