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

Kaan BUYUKDEMIRCI - Bilkent University (Presenter)
Izzet Emre KUCUKKAYA - Technical University of Munich
Eren OLMEZ - Bilkent University
Cagri TORAMAN - Middle East Technical University

MIDDLE EAST TECHNICAL UNIVERSITY

The 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 in the following slides.

- Introduction
  - What is hate speech?
  - The Gap in the Literature
  - Contributions of This Study
  - The Published Dataset

What is hate speech?

Any kind of communication that:

➔Attacks or uses pejorative or discriminatory language

➔To a person or a group

➔On the basis of who they are

The Gap in the Literature

Existing resources mostly:

➔Focus on text sequence classification.

➔Support English.

➔Lacks target detection.

Contributions of This Study

Presenting:

➔A novel dataset

➔A brief summary of related work

➔Benchmark experiments on this novel dataset

The Published Dataset

JL-Hate (Joint Learning Hate Speech Dataset):

➔1,530 tweets, divided equally into English and Turkish

➔Sequence and token classification tasks for joint

- Background
  - Sequence Classification
  - Token Classification
  - Joint Learning
  - Related Work

Sequence Classification

Used to categorize each text as:

➔Hateful

➔Offensive

➔Neutral

Token Classification

Used to identify in text:

➔Signal

➔Target

Joint Learning

Address:

➔Sequence classification

➔Token classification

jointly.

Related Work

				Seq			Tok		
Study	Size	Lng	Domain	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	✓	✓	✓	✓	✓	✓

- Dataset
  - Text-Level Annotations
  - Span-Level Annotations
  - Tokenization and Tagging

Dataset and Annotations

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

51 tweets from each:

➔5 topics

➔3 classes

➔2 languages

Text-Level Annotations

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

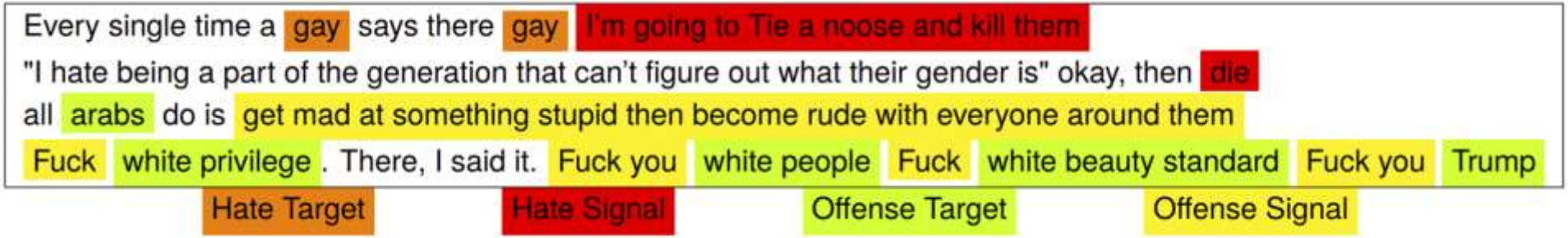
Span-Level Annotations

Definition	EN		TR	
	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

Algorithm 1 Merge Span-Level Annotations

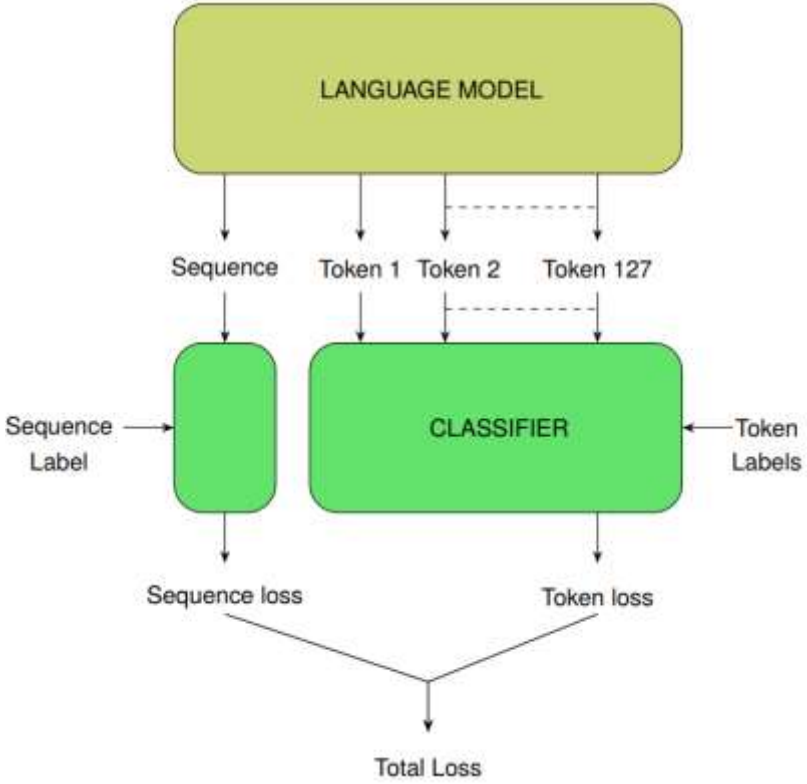
1: Given: $mtla$ (Majority Text-Level Annotation), tla_i (Text-Level Annotation), and sla_i (Span-Level Annotation) for $i \in [1, 4]$
2: Initialize $combined\_span\_annotations$ as an empty list
3: **if** $mtla = S$ **or** $mtla = T$ **then**
4: **return** $combined\_span\_annotations$
5: **end if**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

Tokenization and Tagging



- Experiments
  - Experimental Setup
  - Benchmark Results
  - Error Analysis

Experimental Setup



- ➔10-fold cross-val.
- ➔ConvBERT (TR)
- ➔DistilRoBERTa(EN)
- ➔Split: 0.1 to 0.9
- ➔Weights: 0.1 to 0.9

Benchmark Results

	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

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	-

Error Analysis

"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"

Conclusion

- ➔765 Turkish & 765 English tweets
- ➔Comparison with other studies
- ➔Trained models for each lng.
- ➔Error anal. for both seq. and tok.
- ➔Future work: extend the dataset & relation extraction or detection

Limitations and Ethical Concerns

- ➔ Regulations of social media platforms
- ➔ Human annotation is a costly
- ➔ Relatively smaller size of our dataset

Thank you for listening