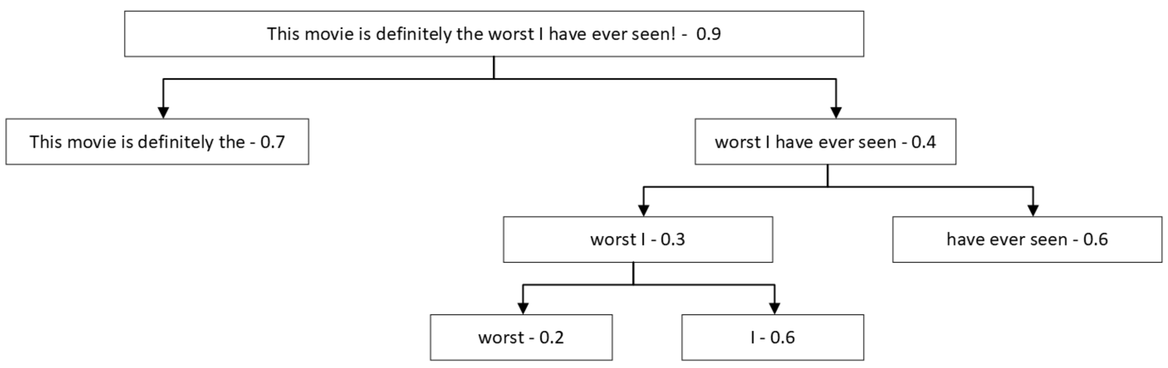
BinarySelect to Improve Accessibility of Black-Box Attack Research

Shatarupa Ghosh and Jonathan Rusert

Purdue University, Fort Wayne

Introduction

- Most **black-box** adversarial attacks use some variation of greedy select to find words to modify.
- Greedy select (and its variations) are **inefficient** since they check every word individually, which leads to at least ***n*** (the length of the text) queries before any modification of the text.
- This inefficiency can be a **barrier to researchers** with fewer resources, as the size of models continues to grow.
- We propose an alternative selection method **BinarySelect** which only requires ***2*log(n)*** queries to find the first word to modify.
- We explore **BinarySelect** in both theoretical performance and apply it in the adversarial attack setting to measure its applied performance against greedy select.

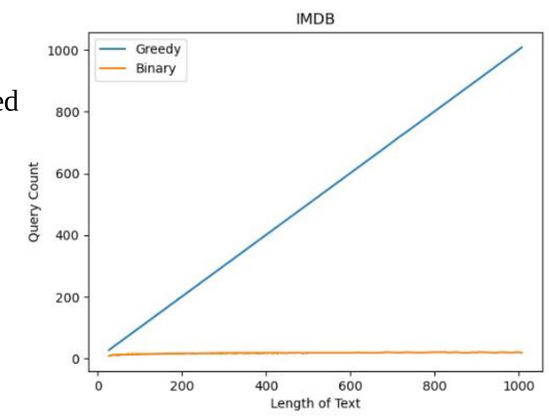


Theoretical Performance

Greedy Select – Always requires ***n*** queries to find any amount of tokens to modify.

BinarySelect

- Best Case: Only 1 word required to be modified
  2*log2(n)
  - Verified on IMDB dataset:



- Average Case: Approx. with BERT-attack, in the worst case (cannot re-use data structure),
  $log2(n) * 2 + log2(n/2) * 2 + log2(n/4) * 2 + ... + log2(n/(2k-1)) * 2$
  - Verified on AG and IMDB:

Token #	AG News	IMDB
1	12.5	17.2
2	17.9	25.0
3	21.6	30.9
4	24.4	35.8
5	26.7	40.0
6	29.0	44.0
7	30.9	47.6
8	32.7	51.0
9	34.5	54.2
10	35.9	57.0
GS	39.5	230.6

- Worst Case: Every word needs to be explored,
  - worse than greedy select

$$n + \sum_{i=1}^{\log_2(n)} n/(2^i)$$

References

- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut.2019. Albert: A lite bert for self-supervised learn-ing of language representations. arXiv preprintarXiv:1909.11942.
- Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue,and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6193–6202, 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 preprint arXiv:1907.11692.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under-standing. 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), pages4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Applied Performance

- We incorporate **BinarySelect** as part of an adversarial attack.
- To compare fairly against greedy select, we use the same word replacement method (WordNet synonyms) for both.

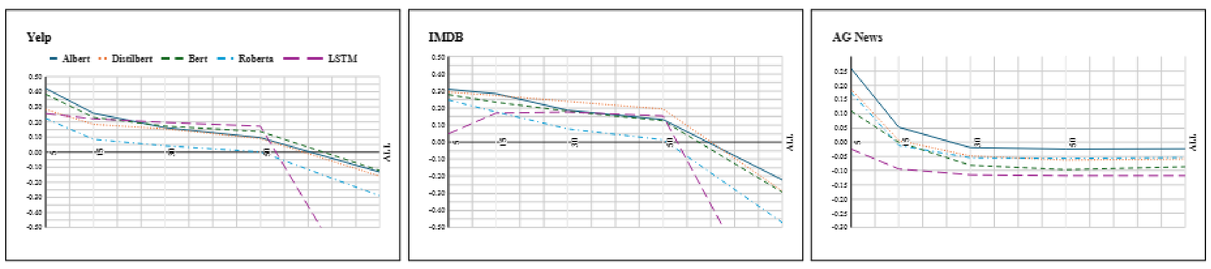
- We test against 5 classifiers (Albert, Distilbert, BERT, RoBERTa, LSTM) across 3 datasets (AG News, Yelp, IMDB) and look at the drop in accuracy compared to the decrease in queries.

- We find a tradeoff, **BinarySelect** reduces the number of queries greatly, with some drop in attack effectiveness.

		Albert		Distilbert		BERT		Roberta		LSTM	
		GS	BS	GS	BS	GS	BS	GS	BS	GS	BS
Yelp	Original Acc.	99.8		95.2		99.5		98.3		94.7	
	Attack Acc.	43.5	51.7	31.1	46.6	47.2	52.6	54.5	65.3	10.9	32.2
	Avg. Queries	217	150	208	141	222	150	239	172	181	119
	Avg. Q's (Success)	156	93	162	93	150	100	160	107	173	91
IMDB	Original Acc.	97.7		96.8		97.9		97.6		84.8	
	Attack Acc.	51.8	66.9	37.2	58.2	54.4	70.0	55.0	72.5	25.4	52.9
	Avg. Queries	318	172	305	156	317	173	332	182	274	136
	Avg. Q's (Success)	273	106	265	99	269	110	275	113	262	96
AG News	Original Acc.	98.8		97.4		99.6		99.2		93.1	
	Attack Acc.	46.2	48.2	60.7	62.8	62.6	64.4	55.9	58.3	43.5	47.7
	Avg. Queries	111	111	121	124	125	127	119	121	104	112
	Avg. Q's (Success)	84	76	92	86	89	84	86	82	84	85

Ablation Studies

- We introduce a variable **k** which restricts how many words the attack can modify and explore how this affects the performance gain of **BinarySelect.**
- We find a greater tradeoff with lower **k**.



- We verify that the tradeoffs hold with **character-level attacks.**

	$k = 5$		$k = 15$		$k = 30$		$k = 50$		$k = ALL$	
	GS	BS	GS	BS	GS	BS	GS	BS	GS	BS
Orig Acc.					85.8					
Attack Acc.	47.8	56.0	29.2	38.5	24.1	32.0	22.8	28.9	22.9	27.9
Avg. Queries	108	31	112	50	117	68	122	85	135	133
Avg. Q's (Success)	101	23	108	34	111	42	113	47	113	49

- We motivate future research by showing an increase in performance when **combining** greedy select and **BinarySelect**.

Model	Attack Acc.	Avg. Q's
GS	3.4	407
BS	3.8	526
Oracle		
$j <= 5$	3.4	369
$j <= 15$	3.4	346
$j <= 30$	3.4	341
$j <= 50$	3.4	358
$j > 50$	3.8	575

Contact Information

Shatarupa Ghosh - shatarupa.ghosh012@gmail.com
Jonathan Rusert – jrusert@pfw.edu

Paper Code - https://github.com/JonRusert/BinarySelect

Paper Presentation - https://youtu.be/9Xmbm9hlBRk

COLING 2025 • Abu Dhabi
The 31st International Conference on Computational Linguistics

On the Robustness of Offensive Language Classifiers

Jonathan Rusert[1], Zubair Shafiq[2], Padmini Srinivasan[1]

[1] University of Iowa, [2] University of California, Davis

Introduction

- Social media platforms are deploying machine learning based offensive language classification systems to combat hateful, racist, and other forms of offensive speech at scale.
- Robustness of offensive classification systems against adversarial attacks has not comphrensively explored.
- We systematically analyze the robustness of state-of-the-art offensive language classifiers against more crafty adversarial attacks that leverage greedy- and attention-based word selection and context-aware embeddings for word replacement.

Threat Model

- Adversary tries to modify their offensive text such that the adversary successfully evades detection, but still preserves semantics and readability for humans.
- For feedback in modifications, the adversary has black-box access to a surrogate classifier (different from the classifier used by the online social media platform).

Offensive Language Classifiers

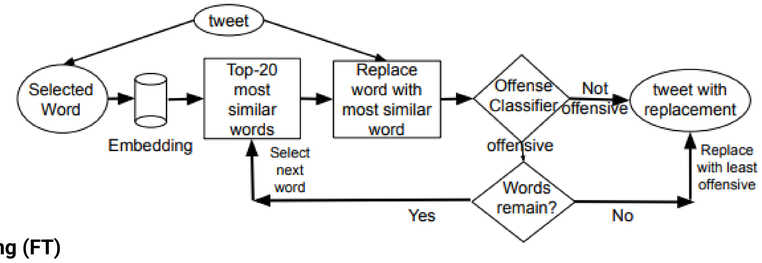
- **NULI** (Liu et al., 2019) – BERT based system trained to identify offensive language. During preprocessing, emojies are converted into English phrases and hashtags are segmented. The top ranked system in OffensEval 2019 (Zampieri et al., 2019)
- **Vradivchev** (Nikolov and Radivchev, 2019) – BERT based system trained on offensive language data. Preprocessing includes removing symbols (“@”, “#”), tokenization, lowercasing, splitting hashtags, removing stopwords. The second best system in OffensEval 2019.
- **MIDAS** (Mahata et al., 2019) – A voting ensemble of a CNN, BLSTM, and BGRU. The top non-BERT system in OffensEval.
- **Offensive Lexicon** (Wiegand et al., 2018) – Simple method that uses a lexicon of offensive words to classify.
- **Perspective API** – Public API which provides a toxicity score for a given text. We use a 0.5 threshold to classify text as offensive. A collaborative creation between Jigsaw and Google.

Proposed Attack (Obfuscation)

Selection

- Greedy Approach (GS) – Remove each word one at a time and calculate drop in classification probability for the text from the surrogate classifier. Remove words until label is flipped. Removed words make up list of possible replacements.
- Attention Approach (AS) – Leverage BLSTM trained on offensive language. Examine attention weights during classification. Select word with highest attention weight to replace. Continue until label flips.

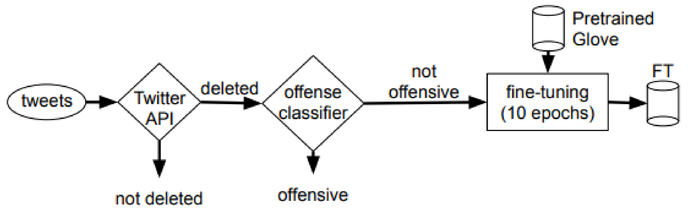
Replacement



Evasion Embedding (FT)

We fine-tune pretrained (Pre) Glove embeddings on the *Evasion* set.

Evasion set created from subset of deleted tweets which were deemed as non-offensive by automatic classifiers.



Main Results

Test on OLID dataset (OffensEval 2019) **(Table shown)**

And SOLID dataset (OffensEval 2020)

Cases where the surrogate classifier is the same as the online classifier are not included as the accuracies easily fall close to 0.

		Drop in Classification Accuracy					
		NULI	vradivchev	MIDAS	Perspective	Lexicon	Avg. Drop
No Attack Accuracy %		61	69	66	68	54	
	Surrogate Classifier						
GS - *Pre*	NULI	-	41	33	34	24	*33*
	vradivchev	28	-	33	28	22	28
	MIDAS	17	35	-	26	19	24
	Perspective	20	36	30	-	17	26
	Average Drop	22	**37**	**32**	**29**	**21**	
GS - *FT*	NULI	-	46	30	31	19	*32*
	vradivchev	39	-	30	26	18	28
	MIDAS	18	29	-	23	13	21
	Perspective	22	37	28	-	13	25
	Average Drop	**26**	**37**	29	27	16	
AS - *Pre*	NULI	-	36	19	19	15	22
	vradivchev	22	-	18	19	17	19
	MIDAS	13	34	-	20	15	21
	Perspective	17	37	23	-	16	*23*
	Average Drop	17	36	20	19	16	
AS - *FT*	NULI	-	39	18	17	15	22
	vradivchev	23	-	17	15	15	18
	MIDAS	11	27	-	17	12	17
	Perspective	17	40	21	-	16	*24*
	Average Drop	17	35	19	16	15	

Table 1: Robustness results on OLID with our attack model. Columns show accuracy drop. The approach is specified as *selection - replacement* where *selection = {Greedy Select (GS), Attention Select (AS)}* and *replacement = {Pre, FT}*. Note that the BLSTM used for AS can be used as an internal classifier but performed poorly so was not included. The adversarial, surrogate classifier is indicated in column 1. The first row presents baseline classification accuracies (%) before attacks. Therefore the resulting accuracies can be calculated by subtracting the drop from the original accuracy.

Human Readability Study

- Crowdworkers annotate attacked text. Take majority vote of 3 crowdworkers per text.

Adversarial Attack	Readability Yes	Partially	No
FT	35	13	2
[%]	70.0	26.0	4.0
Original	37	13	0
[%]	74.0	26.0	0.0
	Conveys same meaning Yes	Partially	No
FT	31	17	2
[%]	62.0	34.0	2.0

Discussion

Comparisons:

- Compare against VIPER (Eger et al. 2019) and Grondahl (Grondahl et al. 2018) character-based attacks.
- We find that unlike the above attacks, the proposed attack is not easily defended against.

FT Embedding Analysis:

- FT embeddings move evasive substitute words closer to offensive probe words.
- Updated embeddings learn creative replacements.

Verifying results:

- We verify results on a Reddit Moderation dataset and find similar outcomes.

Contact Information

Jonathan Rusert – jonathan-rusert@uiowa.edu

Zubair Shafiq – zshafiq@ucdavis.edu

Padmini Srinivasan – padmini-srinivasan@uiowa.edu

References

1. Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In Proceedings of the 13th International Workshop on Semantic Evaluation.
2. Alex Nikolov and Victor Radivchev. 2019. Nikolovradivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In Proceedings of the 13th International Workshop on Semantic Evaluation.
3. Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Ratn Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. Midas at semeval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In SemEval@NAACLHLT.
4. Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
5. Steffen Eger, Gözde Gül ¸Sahin, Andreas Rücklé, JiUng Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. In Proceedings of NAACL-HLT, pages 1634–1647.
6. Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is “love”: Evading hate-speech detection. 11th ACM Workshop on Artificial Intelligence and Security

ACL 2022

Suum Cuique: Studying Bias in Taboo Detection with a Community Perspective

Osama Khalid†, Jonathan Rusert†, Padmini Srinivasan

Introduction

- To identify and mitigate bias, prior research has illustrated the need to consider linguistic norms at the community level when studying taboo (hateful/offensive/toxic etc.) language.
- We propose and test a method to study bias in taboo classification and annotation where a community perspective is front and center.
- This is accomplished by leveraging community language classifiers (CLCs) to represent community level language norms.
- These CLCs help identify bias in both taboo datasets and SOTA taboo classifiers.

Community Language Classifiers (CLCs)

Construction:

- Fine-tune pretrained BERT-base-uncased with a linear layer on top. A softmax function is used to make binary classification on whether an input text belongs to a community or not.
- Models are built with publicly available data from select subreddit communities via Pushshift (Baumgartner et al.).
- Subreddits are grouped into communities based on shared cultural/ethnic heritage determined using subreddit descriptions.

Community	No. of Subreddits	Training set size	Validation set size
Native American (NA)	2	44k	1.4k
Hispanic (HI)	4	95k	6k
Hawaiian (HA)	1	80k	2k
South Asian (SA)	1	101k	6k
African American (AA)	11	70k	5k

Table 1: Dataset details for each community.

Model Validation:

- Verify CLCs on Reddit validation sets.
- Use 0.85 as a threshold to determine whether text is highly aligned (belongs) to a community.

	Reddit Validation Sets				
CLC	NA	HI	HA	SA	AA
NA	**51.8**	1.8	4.5	1.8	2.2
HI	4.3	**58.2**	2.1	2.3	2.2
HA	15.1	6.2	**58.1**	5.1	6.9
SA	6.1	5.2	5.8	**60.7**	20.7
AA	9.8	7.1	8.1	14.4	**64.0**

Table 2: Proportion of each validation set that is highly aligned with each CLC. An alignment score threshold of 0.85 is used to determine high alignment. A text may be aligned with 0 or more models, so column numbers need not sum to 100.

Bias in Taboo Classifiers

Identifying Bias:

- We calculate the Pearson correlation between a CLC's scores and a taboo classifier's scores.
- Use instances which a taboo classifier declared to be taboo.
- Ideally expect a negative correlation – higher taboo classifier confidence mapping to lower community alignment scores and vice versa

Taboo Classifiers Examined:

- **NULI** (Liu et al.) **–** BERT based system trained on offensive language data. Top ranked system at OffensEval.
- **MIDAS** (Mahata et al.) – Ensemble of three deep learning models, CNN, BLSTM, and BGRU trained on offensive language data. The top non-BERT based system at OffensEval.
- **Perspective –** API created by Google and Jigsaw which returns toxicity scores of a given input text.

Results:

- Strong bias found against African American and South Asian communities.
- Correlations with other communities also far from ideal.

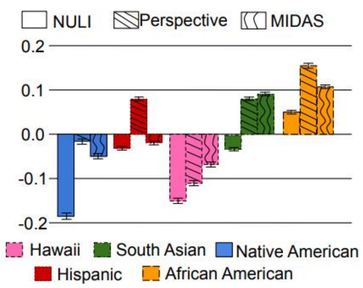


Figure 3: Correlations of taboo classifier scores with community-language classifier scores. Error bars: 95% confidence intervals.

Bias in Taboo Datasets

Identifying Bias:

- Calculate the proportion of taboo labelled texts that are highly aligned with each CLC.
- Expect proportions to be tending towards zero since high alignment means common utterance and thus within the norms.
- Also desire proportions to be even across communities or else a bias will be shown to those communities with higher proportions.

Results:

	Davidson				Gab	Founta		Wiki Toxic		Waseem
CLC	HATE	OFF	OLID	SOLID	Hate	Hate	Abuse	Toxic	Hate	Sexism
NA	14.0	3.9	3.4	1.4	7.6	4.4	3.9	8.0	**13.3**	5.1
HI	5.5	5.2	8.3	3.5	5.5	5.4	6.6	4.9	6.3	3.9
HA	4.3	3.1	6.3	5.1	3.9	6.0	4.6	**9.4**	3.4	4.2
SA	4.2	2.2	**16.3**	5.8	**25.4**	14.5	5.4	8.5	**13.0**	13.9
AA	**20.7**	**29.9**	15.2	**30.4**	12.2	**32.6**	**22.5**	4.9	5.3	**45.5**
Average	9.7	8.9	9.9	9.2	13.2	12.6	8.6	7.1	8.3	14.5
Std. Dev.	7.4	11.8	5.6	11.9	8.7	11.9	7.8	2.1	4.6	17.8

Table 4: Proportion of Taboo datasets with high alignment scores for each CLC. Note, a given text may have high alignment with 0 or more communities. Thus column proportions need not sum to 100.

Small Scale User Study

- Asked 2 African American and 2 South Asian participants to judge their respective texts as offensive/hateful or not.
- Selected texts which had high alignment with the CLCs and high taboo classifier scores.
- both SA annotators disagreed with the classifier assigned taboo labels in 60/78 cases (76.9%) agreeing only in 3/78 comments (3.8%), mixed results in remaining
- AA annotators disagreed with the classifiers for 27/80 (33.8%) comment. They agreed with the classifier's taboo decision 31/80 times (38.8%) and gave mixed judgements in 25%

Future Directions

- Large scale user study to further verify results
- Extend CLCs beyond communities defined by race and ethnicity
- Leverage CLCs to mitigate bias in future taboo classifiers/datasets.

Contact Information

Osama Khalid – osama-khalid@uiowa.edu

Jonathan Rusert – jonathan-rusert@uiowa.edu

Padmini Srinivasan – padmini-srinivasan@uiowa.edu

References

1. Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In Proceedings of the 13th International Workshop on Semantic Evaluation.
2. Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Ratn Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. Midas at semeval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In SemEval@NAACLHLT.
3. Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media, volume 14, pages 830–839
4. Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. ICWSM
5. Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth PortilloWightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. PsyArXiv. July, 18
6. Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the International AAAI Conference on Web and Social Media, volume 12.
7. Zeerak Waseem and Dirk Hovy. 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, San Diego, California. Association for Computational Linguistics.

† denotes equal contribution

ACL 2022

IOWA