Are Chess Discussions Racist? An Adversarial Hate Speech Data Set (Student Abstract)

Rupak Sarkar[1] **and Ashiqur R. KhudaBukhsh**[2]
[1] Maulana Abul Kalam Azad University of Technology
[2] Carnegie Mellon University
rupaksarkar.cs@gmail.com, akhudabu@cs.cmu.edu

Abstract

On June 28, 2020, while presenting a chess podcast on Grandmaster Hikaru Nakamura, Antonio Radić's YouTube handle got blocked because it contained "harmful and dangerous" content. YouTube did not give further specific reason, and the channel got reinstated within 24 hours. However, Radić speculated that given the current political situation, a referral to "black against white", albeit in the context of chess, earned him this temporary ban. In this paper, via a substantial corpus of 681,995 comments, on 8,818 YouTube videos hosted by five highly popular chess-focused YouTube channels, we ask the following research question: *how robust are off-the-shelf hate-speech classifiers to out-of-domain adversarial examples?* We release a data set of 1,000 annotated comments[1] where existing hate speech classifiers misclassified benign chess discussions as hate speech. We conclude with an intriguing analogy result on racial bias with our findings pointing out to the broader challenge of color polysemy.

Introduction

On June 28, 2020, while presenting a chess podcast on Grandmaster Hikaru Nakamura, Antonio Radić's YouTube handle (Agadmator's Chess Channel) got blocked because it contained "harmful and dangerous" content. The channel got reinstated in 24 hours, and YouTube didn't provide any specific reason for this temporary ban. However, Radić speculated that under the current political circumstances[2], a referral to "black against white" in a completely different context of chess, cost him the ban[3]. The swift course-correction by YouTube notwithstanding, in the age of AI monitoring and filtering speech over the internet, this incident raises an important question: *is it possible that current hate speech classifiers may trip over benign chess discussions, misclassifying them as hate speech?* If yes, how often does that happen and is there any general pattern to it? In this paper, via a substantial corpus of 681,995 comments on 8,818 YouTube videos hosted by five highly popular chess-focused YouTube channels, we report our ongoing research on adversarial examples for hate speech detectors.

Hate speech detection, a widely-studied research challenge, seeks to detect communication disparaging a person or a group on the basis of race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby 2000). Hate speech detection research in various social media platforms such as Facebook (Del Vigna et al. 2017), Twitter (Davidson et al. 2017) and YouTube (Dinakar et al. 2012) has received a sustained focus. While the field has made undeniable progress, the domain-sensitivity of hate speech classifiers (Arango, Pérez, and Poblete 2019) and susceptibility to adversarial attacks (Gröndahl et al. 2018) are well-documented.

In this paper, we explore the domain of online chess discussions where *white* and *black* are always adversaries; *kills*, *captures*, *threatens*, and *attacks* each other's pieces; and terms such as *Indian defence*, *Marshall attack* are common occurrences. Our primary contribution is an annotated data set of 1,000 comments verified as **not** hate speech by human annotators that are incorrectly flagged as hate speech by existing classifiers.

Data Set and Hate Speech Classifiers

We consider five chess-focused YouTube channels listed in Table 1. We consider all videos in these channels uploaded on or before July 5, 2020, and use the publicly available YouTube API to obtain comments from these 8,818 videos. Our data set consists of 681,995 comments (denoted by $\mathcal{D}_{chess}$) posted by 172,034 unique users.

We consider two hate speech classifiers: (1) an off-the-shelf hate speech classifier (Davidson et al. 2017) trained on twitter data (denoted by $\mathcal{M}_{twitter}$); and a (2) a `BERT`-based classifier trained on annotated data from a white supremacist

YouTube handles	#Videos	#Comments
Agadmator's Chess Channel	1,780	420,350
MatoJelic	2,976	126,032
Chess.com	2,189	61,472
John Bartholomew	1,619	589,38
GM Benjamin Finegold	254	15,203

Table 1: List of YouTube channels considered.

[1] Available at https://www.cs.cmu.edu/~akhudabu/Chess.html.

[2] https://www.bbc.com/news/world-us-canada-53273381

[3] https://www.thesun.co.uk/news/12007002/chess-champ-youtube-podcast-race-ban/

	$\mathcal{M}_{twitter}$ Davidson et al. 2017	$\mathcal{M}_{stormfront}$ BERT trained on de Gibert et al. 2018
Fraction of positives	1.25%	0.43%
Human evaluation on predicted positives	5% (true positive) 87% (false positive) 8% (indeterminate)	15% (true positive) 82% (false positive) 3% (indeterminate)

Table 2: Classifier performance on $\mathcal{D}_{chess}$.

forum (de Gibert et al. 2018) (denoted as $\mathcal{M}_{stormfront}$).

Results

We run both classifiers on $\mathcal{D}_{chess}$. We observe that only a minuscule fraction of comments are flagged as hate speech by our classifiers (see, Table 2). We next manually annotate randomly sampled 1,000 such comments marked as hate speech by at least one or more classifiers. Overall, we obtain 82.4% false positives, 11.9% true positives, and 5.7% as indeterminate. High false positive rate indicates that without a human in the loop, relying on off-the-shelf classifiers' predictions on chess discussions can be misleading. We next evaluate individual false positive performance by manually annotating 100 randomly sampled comments marked as hate speech by each of the classifiers. We find that $\mathcal{M}_{twitter}$ has slightly higher false positive rate than $\mathcal{M}_{stormfront}$. Also, $\mathcal{M}_{stormfront}$ flags substantially fewer comments as hate speech than $\mathcal{M}_{twitter}$. Since $\mathcal{M}_{stormfront}$ is trained on a white supremacist forum data set, perhaps this classifier has seen hate speech targeted at the black community on a wider range of contexts. Hence, corroborating to the well-documented domain-sensitivity of hate speech classifiers, $\mathcal{M}_{stormfront}$ performs slightly better than $\mathcal{M}_{twitter}$ trained on a more general hate speech twitter data set. Table 3 lists a random sample of false positives from $\mathcal{M}_{stormfront}$ and $\mathcal{M}_{twitter}$. We note that presence of words such as `black`, `white`, `attack`, `threat` possibly triggers the classifiers.

That is one of the most beautiful attacking sequences I have ever seen, black was always on the back foot. Thank you for sharing. Seeing your channel one day in my recommended got me playing chess again after 15 years. All the best.
At 7:15 of the video Agadmator shows what happens when Black goes for the queen. While this may be the most interesting move, the strongest continuation for Black is Kg4. as Agadmator states, White is still winning. But Black can prolong the agony for quite a while.
White's attack on Black is brutal. White is stomping all over Black's defenses. The Black King is gonna fall...
That end games looks like a draw to me... its hard to see how it's winning for white. I seems like black should be able to block whites advance.
...he can still put up a fight (i dont see any immediate threat black can give white as long as white can hold on to the e-rook)

Table 3: Random samples of misclassified hate speech.

Black is to Slave as White is to?

We conclude our paper with a simple yet powerful observation. Word associations test (e.g., *France*:*Paris*::*Italy*:*Rome*) on Skip-gram embedding spaces (Mikolov et al. 2013) are well-studied. Social biases in word embedding spaces is a well-established phenomenon (Garg et al. 2018) with several recent lines of work channelised to debiasing efforts (Manzini et al. 2019). We perform the following analogy test: *black*:*slave*::*white*:*?*, on $\mathcal{D}_{chess}$ and two data sets containing user discussions on YouTube videos posted on official channels of Fox News ($\mathcal{D}_{fox}$) and CNN ($\mathcal{D}_{cnn}$) in 2020. While both $\mathcal{D}_{fox}$ and $\mathcal{D}_{cnn}$ predict *slavemaster*, $\mathcal{D}_{chess}$ predicts *slave*! Hence, the *captures*, *fights*, *tortures* and *killings* notwithstanding, over the 64 black and white squares, the two colors attain a rare equality the rest of the world is yet to see.

References

Arango, A.; Pérez, J.; and Poblete, B. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *SIGIR*, 45–54.

Davidson, T.; Warmsley, D.; Macy, M. W.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM 2017*, 512–515.

de Gibert, O.; Perez, N.; García-Pablos, A.; and Cuadros, M. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 11–20.

Del Vigna, F.; Cimino, A.; Dell'Orletta, F.; Petrocchi, M.; and Tesconi, M. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 86–95.

Dinakar, K.; Jones, B.; Havasi, C.; Lieberman, H.; and Picard, R. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3): 18.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115(16): E3635–E3644.

Gröndahl, T.; Pajola, L.; Juuti, M.; Conti, M.; and Asokan, N. 2018. All You Need is" Love" Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12.

Manzini, T.; Yao Chong, L.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *NAACL-HLT*, 615–621.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Nockleby, J. T. 2000. Hate speech. *Encyclopedia of the American constitution* 3(2): 1277–1279.