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

**Rupak Sarkar**
Maulana Abul Kalam Azad University of Technology

**Ashiqur R. KhudaBukhsh**
Carnegie Mellon University

## 1 A Real-world Ban

- On 28th June 2020, while streaming a chess podcast of Grandmaster Hikaru Nakamura, Agadmator, a highly popular YouTuber, received a temporary streaming ban


- His channel got reinstated in 24 hours
- YouTube did not provide any explanation

Image courtesy: followchess.com

## 2 What Could Have Happened?

- We don't exactly know the inner workings of big-tech speech filtering systems

- We investigate this ban with previously published
  - Methods: [Davidson et al.; ICWSM 2017], [Devlin et al.; NAACL 2019]
  - Data: [Davidson et al.; ICWSM 2017], [de Gilbert et al.; AWL2 2018]

## 3 Research Question and Methodology

- Research question: Is it possible that current hate speech classifiers may trip over benign chess discussions, misclassifying them as hate speech?
- Method: Existing hate speech classifiers trained on previously published data sets are run on a new data set of YouTube chess discussions
  - 5 prominent chess YouTubers
  - More than 8K videos
  - More than 600K comments on these videos

## 4 Our Results

- Yes, existing hate speech classifiers do misclassify chess discussions as hate speech
- More than 80% comments flagged as hate speech are benign chess discussions

| | $\mathcal{M}_{twitter}$ 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)<br>87% (false positive)<br>8% (indeterminate) | 15% (true positive)<br>82% (false positive)<br>3% (indeterminate) |

Classifier performance on chess discussions

## 5 An Adversarial Hate Speech Data Set

| |
|---|
| *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)* |

- An adversarial data set of 1,000 challenging examples from chess[1]

1. Available at cs.cmu.edu/~akhudabu/Chess.html

## 6 Insights

- Presence of words like "black," "white," "attack," and "threat" triggers the classifiers.
- Classifier trained on a data set from a white supremacist forum makes fewer mistakes

## 7 Black : slave :: white :?

- Word analogy tests are a powerful technique to uncover social bias [Manzini et al.; NAACL 2019]
- black : slave :: white : slavemaster
- black : slave :: white : slave
- Over the 64 black and white squares, the two colors attain a rare equality the rest of the world is yet to see

## 8 Conclusion

- A compelling case-study of domain-sensitivity of hate speech classifiers inspired from a real-world ban
- A novel, annotated adversarial *hate speech* data set
- Broader questions on
  - domain-sensitivity of content classifiers
  - color polysemy
  - potential risks of reliance on AI without human-in-the-loop