Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech

Antonios Anagnostou, Ioannis Mollas, Grigorios Tsoumakas
School of Informatics, Aristotle University of Thessaloniki
{anagnoad,iamollas,greg}@csd.auth.gr

Abstract

Hatebusters is a web application for actively reporting YouTube hate speech, aiming to establish an online community of volunteer citizens. Hatebusters searches YouTube for videos with potentially hateful comments, scores their comments with a classifier trained on human-annotated data and presents users those comments with the highest probability of being hate speech. It also employs gamification elements, such as achievements and leaderboards, to drive user engagement.

1 Introduction

Hate speech can be defined as speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability or gender[1]. Hateful speech is widespread in the public debate, including on online platforms, social media and forums [Saleem *et al.*, 2016].

Online hate speech has debilitating consequences on individual victims' well-being by imposing psychological harm, damaging self-worth and inducing fear. The duration of the exposure maintained by the online availability of the content is associated with greater damage on victims and greater empowerment of perpetrators compared to offline hate speech [Gagliardone *et al.*, 2015]. Therefore, the sooner the content is down, the better the chances to mitigate the negative effects of hate speech on victims' well-being.

The way social media companies manage hate speech is to rely on users to report it, after which they apply manpower to review and take down the relevant content. Towards automating this process, they are working on AI techniques, but it is hard for computers to understand the differences between a truly offensive post and something posted by a well-meaning critic or by a satirist [Murgia and Warrell, 2017]. Manually reviewing all uploaded content is humanly impossible due to the scale of the content.

In 2016, EU together with Facebook, Microsoft, Twitter and YouTube (the *IT Companies*) presented a *Code of conduct on countering illegal hate speech online*[2] to formalize the reporting of hate speech. Among their main commitments, IT Companies agreed to review the majority of valid notifications for removal of illegal hate speech in less than 24 hours and remove or disable access to such content, if necessary.

However, by itself this is not enough to reduce the spreading of illegal online hate speech. What is actually needed is an increase in the requests for hate speech removal. This can be achieved by building tools that make it easy for ordinary citizens to report illegal online hate speech. We here present such a tool, dubbed Hatebusters[3], aiming to establish an online community of volunteer citizens actively reporting illegal online hate speech in YouTube. Hatebusters searches YouTube for videos with potentially hateful comments, scores their comments with a machine learning classifier trained on human-annotated data and presents users those comments with the highest probability of being hate speech. Hatebusters employs gamification elements (achievements and leaderboards) to increase user engagement.

2 Related Work

A number of applications for online reporting of hate speech already exist. The hate speech reporting portal[4] of the EU project **MANDOLA** (Monitoring and Detecting OnLine Hate Speech) points citizens to European organizations, where they can report hate speech. **Hate Speech Watch**[5] is a user-generated repository to trace, share and discuss online hate speech content. Instances of hate speech signaled there are not reported to judicial authorities, regulatory bodies or Internet providers. Project **CONTACT** (Creating an Online Network, monitoring Team and phone App to Counter hate crime Tactics) offers a web form for reporting hate speech, as a well as a live dashboard presenting statistics of hate speech reporting[6]. Project **eMORE**[7] (Monitoring and Reporting Online Hate Speech in Europe) is building a crawler to monitor hate speech and an app to report hate speech. Finally, the EU funded project **INACH** (International Network Against Cyber Hate) offers an email address for reporting complaints, which are then dispatched to appropriate organizations[8].

[1] https://en.wikipedia.org/wiki/Hate_speech
[2] http://europa.eu/rapid/press-release_IP-16-1937_en.htm
[3] https://hatebusters.org
[4] http://mandola-project.eu/portal/report.html
[5] https://www.nohatespeechmovement.org/hate-speech-watch
[6] http://reporthate.eu
[7] http://www.emoreproject.eu/about-project
[8] http://www.inach.net/index.php

All of the above take a *passive* approach to online hate speech reporting, in contrast to Hatebusters. Users typically visit such sites after they run into hate speech, during their normal online activities. In addition, such applications lack elements of user engagement and community building. Finally, they are not connected to the IT companies and therefore cannot contribute directly to the implementation of the code of conduct.

Hate speech detection has recently gained a lot of attention from the NLP and ML communities. It is a very difficult task, where keyword-based approaches fall short [Gitari *et al.*, 2015; Saleem *et al.*, 2016; Davidson *et al.*, 2017; Waseem *et al.*, 2017; Nobata *et al.*, 2016]. Low quality of annotations is an important issue [Ross *et al.*, 2016]. Hate speech could be no exception to the rule of neural network models [Badjatiya *et al.*, 2017; Djuric *et al.*, 2015]. Most of the proposed approaches focus on Twitter [Park and Fung, 2017; Badjatiya *et al.*, 2017; Davidson *et al.*, 2017; Burnap and Williams,], while some on news agencies [Gao and Huang, 2017; Nobata *et al.*, 2016]. On the other hand, Hatebusters is the only work to focus on YouTube comments, and follows a straightforward supervised learning approach, as building a robust working system was our initial priority.

3 Hatebusters

Hatebusters is a synthesis of three components: a hate speech classification engine, a backend service with runtime and time-scheduled tasks and a user-centric platform exposing candidate hate speech comments.

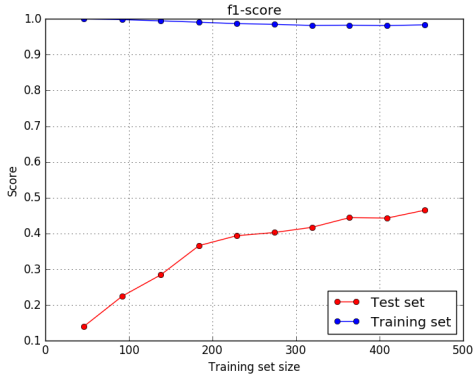
3.1 Hate Speech Classification Engine

Hatebusters' hate speech classification engine, dubbed *Hatewise*, uses a Support Vector Machine for the classification of YouTube comments as hateful or not. We represent each comment using the bag-of-words model and compute tf-idf values for each word. For each comment, the classifier outputs a confidence score, which is effectively the distance from the classifier's hyperplane. Hatebusters utilizes this confidence score in two manners. For platform administrators, we expose comments which are closer to the hyperplane, following the *uncertainty sampling* technique [Lewis, 1995] for active learning. This way we aim to disambiguate the comments for which the classifier is the least certain. For all other members of the community, we expose the comments for which the classifier is most confident about.

The first classifier of Hatewise was trained in March 2017 on a training set that comprised 454 samples, 334 negative and 120 positive ones. To collect this set of samples, we used the Hatebusters pipeline and exposed retrieved comments to the administrators of the platform (i.e. ourselves) without any ranking. Our aim at that stage was tagging the samples rather than reporting them to YouTube. The learning curve for this data set, shown in Figure 1, hinted to us that acquiring further samples would increase prediction accuracy.

After the initial announcement of Hatebusters to external users and members of our laboratory, the data set was expanded to a total of 936 samples, out of which 358 were positive ones. We trained a second classifier in April 2017 using this extended data set. F1-score increased from 0.459 to 0.654 and average precision from 0.556 to 0.722.

Figure 1: Learning curve showing the F1-score of our classifier for different sizes of the training set.



3.2 Backend and Frontend Services

The backend provides both run-time services as well as time-scheduled tasks. The *run-time services* include mainly *user-centric APIs* used to register, authenticate, save and obtain details about users as well as *comment-centric APIs* which are mainly *time-scheduled tasks* operations such as:

- Obtaining new comments daily by querying the YouTube Data API. Hatebusters stores the comment IDs, the comments' text as well as other metadata (video id, title etc) and it pairs them with a hate speech score generated by the Hatewise engine.
- Checking for deleted comments; these can be the result of an author retracting their comment, a deleted parent-comment or the hosting video being deleted altogether. We store the deleted comments as a new entry in the database so as to keep the videos metadata for usage during classification experiments.

The frontend of Hatebusters is designed to be engaging and to provide stimuli to users to report an increasing amount of YouTube comments. To achieve this goal, we used gamification elements such as a *point system*, which ranks the user among their peers and *achievements* which are given to users so as to engage them to play more and report more comments.

4 Future Work

Hatebusters is currently being revised towards version 2. For improving our classification engine, we are working on different representations of comments, results diversification [Abid *et al.*, 2016], utilizing deleted comments and considering user information, as well as on obtaining a new quality-controlled humanly annotated sample of data, thanks to (co)winning CrowdFlower's AI for Everyone Challenge for Q4 of 2017[9]. For improving user engagement, we are working towards introducing more achievements and weekly digest emails.

[9]https://www.crowdflower.com/announcing-q4-ai-everyone-winners/

References

[Abid *et al.*, 2016] Adnan Abid, Naveed Hussain, Kamran Abid, Farooq Ahmad, Muhammad Shoaib Farooq, Uzma Farooq, Sher Afzal Khan, Yaser Daanial Khan, Muhammad Azhar Naeem, and Nabeel Sabir. A survey on search results diversification techniques. *Neural Computing and Applications*, 27(5):1207–1229, 2016.

[Badjatiya *et al.*, 2017] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[Burnap and Williams,] Pete Burnap and Matthew L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[Djuric *et al.*, 2015] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30. ACM, 2015.

[Gagliardone *et al.*, 2015] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering Online Hate Speech*. UNESCO Series on Internet Freedom. 2015.

[Gao and Huang, 2017] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266. INCOMA Ltd., 2017.

[Gitari *et al.*, 2015] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[Lewis, 1995] David D. Lewis. A sequential algorithm for training text classifiers. *ACM SIGIR Forum*, 29(2):13–19, 1995.

[Murgia and Warrell, 2017] Madhumita Murgia and Helen Warrell. Why tech companies struggle with hate speech. *Financial Times*, 2017.

[Nobata *et al.*, 2016] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[Park and Fung, 2017] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45. Association for Computational Linguistics, 2017.

[Ross *et al.*, 2016] Björn Ross, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In Michael Beißwenger, Michael Wojatzki, and Torsten Zesch, editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9, 2016.

[Saleem *et al.*, 2016] Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. In *In proceedings of the 1st workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS), collocated with LREC 2016*, 2016.

[Waseem *et al.*, 2017] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.