Detecting harassment in real time as conversations develop

Wessel Stoop
CLST, Radboud University
w.stoop@let.ru.nl

Florian Kunneman
CLST, Radboud University
f.kunneman@let.ru.nl

Antal van den Bosch
KNAW Meertens Instituut
antal.van.den.bosch@meertens.knaw.nl

Ben Miller
Emory University
b.j.miller@emory.edu

Abstract

We developed a machine-learning-based method to detect video game players that harass teammates or opponents in chat earlier in the conversation. This real-time technology would allow gaming companies to intervene during games, such as issue warnings or muting or banning a player. In a proof-of-concept experiment on *League of Legends* data we compute and visualize evaluation metrics for a machine learning classifier as conversations unfold, and observe that the optimal precision and recall of detecting toxic players at each moment in the conversation depends on the confidence threshold of the classifier: the threshold should start low, and increase as the conversation unfolds. How fast this *sliding threshold* should increase depends on the training set size.

1 Introduction

In many online platforms that allow user interaction, verbal harassment has become commonplace. For example, a survey by The Wikimedia Foundation showed that '38% of the 3,845 Wikimedia editors that were surveyed (an estimated total over 130,000) had experienced some form of harassment, and over half of those contributors felt a decrease in their motivation to contribute in the future' (Wulczyn et al., 2017). In this work we would like to focus on harassment in the online gaming community, where so-called *toxic players* are the subject of frequent media attention. For some video games over 1% of the player base is estimated to be consistently toxic[1]. Yet, for the game *League of Legends*, researchers found that this 1% of the player population only accounted for 5% of the toxic speech. The former director of Riot Games' Player Behavior Unit attributes most toxicity to "the average person just having a bad day" (Maher, 2016). As encounters with harassment are a major predictor for players quitting a video game[2], creating healthy communities is an important focus point for many video game developers[3].

There has been an increase recently in the number of academic papers on automatically detecting harassment; see Zhang et al. (2018b) and van Aken et al. (2018) for overviews. Many of these works focus on datasets with relatively short conversations (often <20 turns), consisting of longer utterances (often multiple full sentences). As a result, most of these studies approach detecting verbal harassment as a classical text classification task, where each individual comment is considered a document on its own that should be assigned one of two or more categories. Conversations in video games, on the other hand, are different in nature: they consist of up to several hundreds of utterances, depending on the length of a match in the chosen video game, and these utterances are usually shorter, at least partly due to the restriction that the act of typing temporarily prevents players from playing. For this reason, we focus less on rating individual *comments* (an individual swear word or insult does not indicate harassment per se), but instead on detecting *players* within a match that consistently and knowingly harass teammates and/or opponents.

Self-policing of communities has been implemented by many game companies, among other things in the form of post-game ratings by other players. Based on this information, video game developers already have a good estimate of which players behaved badly at what time, so an au-

[1] https://www.youtube.com/watch?v=HQwL6zh7AgA&feature=youtu.be&t=39m38s

[2] https://www.youtube.com/watch?v=HQwL6zh7AgA&feature=youtu.be&t=33m57s

[3] https://kotaku.com/league-of-legends-neverending-war-on-toxic-behavior-1636894289

Proceedings of the Third Workshop on Abusive Language Online, pages 19–24
Florence, Italy, August 1, 2019.

tomated system that makes this estimate retroactively would not be of much added value. Instead, toxic players should be detected as the conversation develops, as early as possible, making it possible for gaming companies to intervene in one way or the other (like warning, muting or banning a player). Translated to a machine learning task, this means that instances (e.g.: players) change over time, as more information about the instances (more utterances) becomes available. This leads to *time* as an extra dimension of interest for metrics like precision, recall and F-score: instead of presenting them as a single number, it should be represented how they change during the conversation.

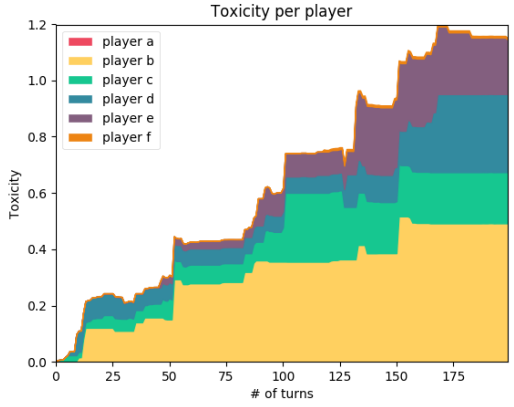

Figure 1: Classifier confidence for the 'toxic' class for six players during a single conversation.

A visualization of the estimated 'temperature' of a single conversation over time is given in Figure 1. In this work we will apply this idea of detecting harassment over the course of a conversation at scale, to evaluate various (parameters of) classifiers during the course of a conversation. More specifically, we will show that the optimal confidence threshold above which a player can be considered toxic increases as a conversation evolves, and that the rate of this increase interacts with the amount of training material.

2 Related work

The task of harassment detection in online conversation relates to tasks like cyberbullying and hate speech detection (van Aken et al., 2018). Despite differences in terminology and definitions of these terms, similar methods can often be applied; we will therefore treat it as one research field.

Early approaches to detecting harassment employ a simple lexicon or 'classic' machine learning algorithms such as Support Vector Machines, Naive Bayes, Logistic Regression, and Random Forests (see Schmidt and Wiegand (2017) for an overview) and focus on manually extracted features. Besides word or character n-grams and POS tags, the approaches typically make use of features such as punctuation, word and document length, capitalization, and gender identity of the speaker (Davidson et al., 2017; Nobata et al., 2016; Waseem, 2016; Waseem and Hovy, 2016). Many of these approaches have the advantage of explainability (to a certain extent), but struggle when harassment is implicit (Dinakar et al., 2011) or when harassment-related words have multiple meanings (Kwok and Wang, 2013; Davidson et al., 2017).

Some works apply these techniques to harassment in video games specifically: lexicon-based approaches have been shown to be useful for the games *DotA* (Märtens et al., 2015), *StarCraft II* (Thompson et al., 2017) and *World of Tanks* (Murnion et al., 2018), whereas Balci and Salah (2015) apply a Bayesian Point Machine to the game *Okey*. Of particular relevance is the study by Blackburn and Kwak (2014), who use the crowd sourced Tribunal decisions in the game *League of Legends* as their ground truth, similar to this paper (see Section 3). Besides language data, they feed a Random Forest classifier with various game-specific features, such as the number of kills and deaths, and the type of report by other players. The combined model can emulate Tribunal decisions with an Area Under the ROC Curve (AUC) of 80%.

More recent studies often use deep neural networks, with the most popular architectures being Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The main advantage of the former is its ability to extract useful features, while the latter is well suited for the sequential nature of language. Zhang et al. (2018b) conduct an extensive evaluation of approaches for detecting hate speech so far and propose a combination of CNNs and RNNs to outperform them. Similarly, van Aken et al. (2018) do an in-depth error analysis for various approaches to toxic comment classification, and propose an ensemble method to outperform them.

Whereas most of these studies classify individual utterances, there are also works with a broader scope. Focusing on users instead of utterances,

Cheng et al. (2015) aim to detect 'antisocial users' in online communities over a longer period of time. They observe that the post quality of users labeled as antisocial worsens over time, possibly related to being censored. Using a variety of features as input, they use logistic regression to predict which users will be banned in the future. They achieve an AUC of 80% after observing 5–10 posts. Focusing on early instead of retrospective detection, Zhang et al. (2018a) try to predict whether the relatively short conversations on Wikipedia talk pages (average 4.6 utterances) will derail based on the first few utterances. While humans can do this with 72% accuracy, their 'Perspective API' achieves a score of 64.9%.

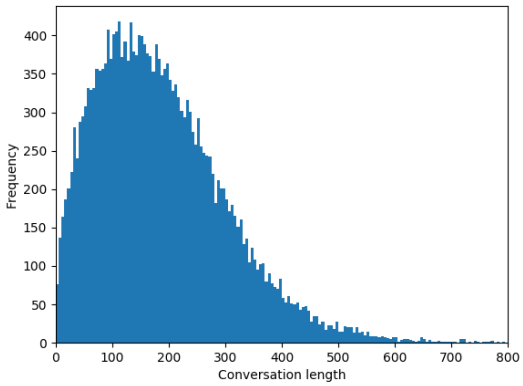

Figure 2: The number of utterances per conversation in our dataset.

3 Dataset

As a dataset, we use 5000 conversations from the video game *League of Legends*, obtained from video game developer Riot Games, containing utterances by 48512 players. Toxic players in this dataset were first identified by team mates and opponents, and later reassessed by other members of the community in a voting system called the 'Tribunal'. Only cases where a so-called 'overwhelming majority' was reached were considered toxic.

An average conversation in our dataset consists of 186.77 utterances (standard deviation 122.01), as visualized in figure 2, by 9.7 speakers (standard deviation 6.07). An average utterance consists of 3.15 words (standard deviation 2.63). 10.3% of the speakers in our dataset were labeled toxic by the Tribunal.

A typical case of harassment looks like this:

```
Z fukin bot n this team....
  so cluelesss gdam
V u cunt
A WTF
J TSM
V TSMMM
A 35 baron
Z wow voli....u jus let them kill
  me....instead of peeling
V ARE YOU RETARDED
L cheesed?
V U ULTED INTO 4 PEOPLE
D no death rocket plz
V HOW DO I PEED FOR UR AUTISTIC
  ASS
V ur mom should have swallowed you
Z this game is like playign with
  pre 30s lol....complete
  clueless lewl
L ur shyt zed
V AUTISM
D Oh bby|
```

Pilot experiments showed that the three main predictors for toxicity in this dataset are swear words, insults and talking about losing, all of which are present in this example ('fukin', 'u cunt', 'u jus let them kill me', respectively).

4 Method

To monitor conversations in progress and evaluate the success, we developed the framework HaRe (Harassment Recognizer)[4]. During a conversation, HaRe keeps track of toxicity estimates for all participants separately, updating the estimate for each speaker every time s/he makes an utterance. This is done by concatenating all utterances for that speaker, separated by *[NEW UTTERANCE]* tags, and classifying the resulting text. As an example, to obtain toxicity estimates in a conversation where three players each have generated six utterances so far, this means the classifier is asked to classify three texts, all containing five *[NEW UTTERANCE]* tags. All graphs in this work were created by the HaRe visualization module.

For classifier setup, we adopted the best performing neural network architecture in the *Toxic Comment Classification Challenge* on Kaggle[5], feeding a sequence of words to an RNN with an embedding layer (300 dimensions), two bidirectional GRU layers (16 units) feeding into two final dense layers (256 units). The output layer is a single sigmoid unit indicating the network's confidence that the input text is toxic. This is imple-

[4]The software and source code for HaRe is available at https://github.com/woseseltops/HaRe

[5]The setup is explained here: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557

mented in HaRe and uses TensorFlow under the hood (Abadi et al., 2015).

We split the dataset into 1000 conversations for evaluation and 4000 for training (but in figure 6 we also experiment with smaller training set sizes). Training texts were created by concatenating all utterances per player, similar to how conversations are offered to the classifier during the classification phase. Important differences between the training and classification phase are (1) the texts in the training phase were downsampled to have an equal 50%-50% distribution of toxic and non-toxic texts, while during the classification phase only 10.3% of the texts were labeled toxic, and (2) training was done on full conversations that had finished, while during the classification phase the conversations were most often not finished yet (so the texts to classify in the beginning of conversations were considerably shorter).

5 Results

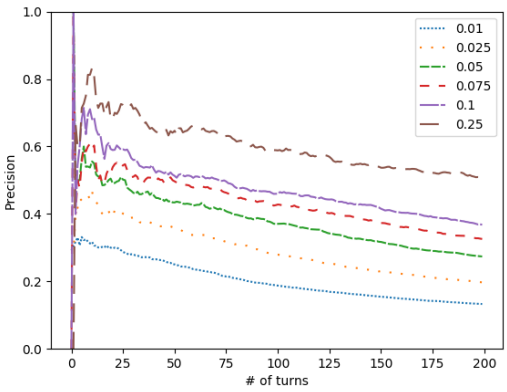

Figure 3: Precision recognizing toxic players over the course of conversations for various confidence thresholds.

Figures 3, 4 and 5 visualize the precision, recall and F-score of our classifier as the conversation unfolds, aggregated over our 1000 test conversations. They were created using a classifier trained on 4000 conversations and various thresholds. We see recall increase during a conversation as more information on each of the players (that is, more utterances) becomes available. However, every new utterance is also an extra source of information that could incorrectly be interpreted as an indicator for toxicity, leading to a decrease in precision during a conversation.

The rate of the recall increase and precision decrease over time greatly depend on the confidence level above which a player is considered toxic. Interestingly, this leads to a situation where the optimal threshold (that is, the threshold that results in the highest F-score) changes over the course of a conversation: whereas in the beginning the threshold should be as low as possible, it should generally be increased as the conversation progresses and more data to work with (more utterances) becomes available.

Figure 6 shows the results of retroactively selecting the threshold with the highest F-score for each turn in the conversation, for classifiers trained on various amounts of data. We observe that the rate in which this sliding threshold should be increased itself depends on the size of the training set: the larger the training set, the slower the threshold can be increased.

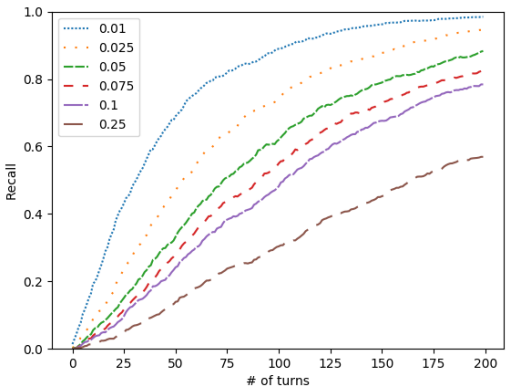

Figure 4: Recall recognizing toxic players over the course of conversations for various confidence thresholds.

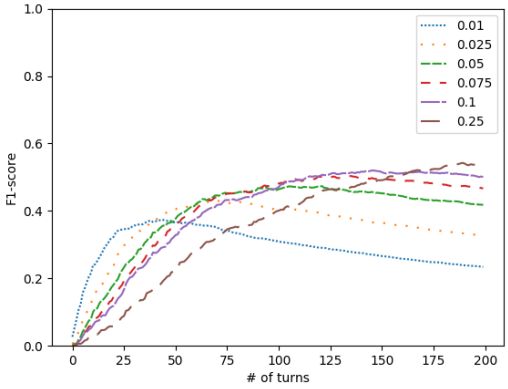

Figure 5: F-score recognizing toxic players over the course of conversations for various confidence thresholds.

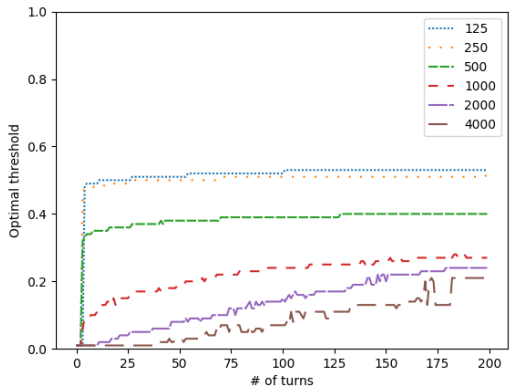

Figure 6: Confidence thresholds for optimal F-scores over the course of conversations for various training set sizes.

6 Discussion and conclusion

In this work we focused on detecting harassment as early as possible in a video game chat session, and observed that the classifier confidence threshold should start low and should be moved up during a conversation as more material for each speaker becomes available, for an optimal F-score at each point in the conversation. The exact starting point and rate of increase of this *sliding threshold* of course depend on the classifier setup and dataset; we showed for example that there seems to be an interaction with the training set size. To decide the optimal values for these two parameters for conversation monitoring software, creating a graph like figure 6 could be useful.

A downside of the approach presented here is that low recall scores are ambiguous in interpretation: they could either indicate a badly performing classifier missing actual harassment, or a lack of harassment so far. For both reasons evaluation measures tend to be low in the first few turns of a conversation. Furthermore, all evaluation metrics used focus on toxicity and ignore whether the classifier is making correct negative judgements at any point; this would call for metrics such as Area Under the ROC Curve.

Our approach should be compared to an approach that labels harassment at the utterance level. This may help pinpoint the exact moment at which the toxic player started using toxic language; this may be earlier than the point at which our confidence threshold is exceeded.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 33–42, Brussels, Belgium. Association for Computational Linguistics.

Koray Balci and Albert Ali Salah. 2015. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Computers in Human Behavior*, 53:517–526.

Jeremy Blackburn and Haewoon Kwak. 2014. Stfu noob!: Predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 877–888, New York, NY, USA. ACM.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. *CoRR*, abs/1504.00680.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. 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.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Papers from the 2011 ICWSM Workshop*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pages 1621–1622. AAAI Press.

Brendan Maher. 2016. Can a video game company tame toxic behaviour? *Nature News*, 531(7596):568.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6. IEEE.

Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. 2018. Machine learning and semantic analysis of in-game chat for cyberbullying. *Computers & Security*, 76:197–213.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. pages 1–10.

Joseph J Thompson, Betty HM Leung, Mark R Blair, and Maite Taboada. 2017. Sentiment analysis of player chat messaging in the video game starcraft 2: Extending a lexicon-based model. *Knowledge-Based Systems*, 137:149–162.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

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.

Ellery Wulczyn, Dario Taraborelli, Nithum Thain, and Lucas Dixon. 2017. Algorithms and insults: Scaling up our understanding of harassment on wikipedia. https://blog.wikimedia.org/2017/02/07/scaling-understanding-of-harassment/. Accessed: 2019-04-11.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018a. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018b. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web*, pages 745–760, Cham. Springer International Publishing.