Technology Solutions to Combat Online Harassment

George W. Kennedy Intel
Andrew McCollough EdgeRock Tech Partners
Edward Dixon Intel
Alexei Bastidas Intel
John Ryan Intel
Chris Loo Intel
Saurav Sahay Intel
orge.w.kennedy@intel.com ward.dixon@intel.com

Abstract

This work is part of a new initiative to use machine learning to identify online harassment in social media and comment streams. Online harassment goes underreported due to the reliance on humans to identify and report harassment, reporting that is further slowed by requirements to fill out forms providing context. In addition, the time for moderators to respond and apply human judgment can take days, but response times in terms of minutes are needed in the online context. Though some of the major social media companies have been doing proprietary work in automating the detection of harassment, there are few tools available for use by the public. In addition, the amount of labeled online harassment data and availability of cross platform online harassment datasets is limited. We present the methodology used to create a harassment dataset and classifier and the dataset used to help the system learn what harassment looks like.

1 Introduction

Online harassment has been a problem to a greater or lesser extent since the early days of the internet. Previous work has applied anti-spam techniques like machine-learning based text classification (Reynolds, Kontostathis, & Edwards, 2011) to detecting harassing messages. However, existing public datasets are limited in size, with labels of varying quality.

The #HackHarassment (#HackHarassment, n.d.) initiative (an alliance of tech companies and NGOs devoted to fighting bullying on the internet) has begun to address this issue by creating a web tool to collect and label data, and using the tool to generate a large, high-quality, cross-platform dataset. The release of this tool is scheduled for June 2017. As we complete further rounds of labelling with a public audience, later iterations of this dataset will increase the available samples by at least an order of magnitude and enable corresponding improvements in the quality of machine learning models we have built for harassment detection. In this paper, we introduce an improved cross-platform harassment dataset and a machine learning model built on the dataset.

2 Related Work

Previous work in the area by (Bayzick, Kontostathis, & Edwards, 2011) showed that natural language processing in combination with a rule-based system could detect bullying messages on an online forum, but with very poor accuracy. However, the same work also made clear that the limiting factor on such models was the availability of a suitable quantity of labeled examples, e.g. the

Bayzick work relied on a dataset of 2,696 samples, only 196 of which were found to be examples of bullying behavior. Additionally, this work relied on classical decision-tree models like J48 and JRIP, and k-nearest neighbors classifiers like IBk, as opposed to modern ensemble methods or deep neural-network-based approaches. In addition, Intel's #HackHarassment team published work (Bastidas et al., 2016) showing results for harassment detection using a variety of model types on a new dataset of comments and posts which their team had labelled.

More recently, major internet companies have focused efforts on combating various forms of harassment online. Yahoo researchers have developed machine learning models for detecting abusive language (Nobata, Tetreault, Thomas, Mehdad, & Chang, 2016a) and a Google Jigsaw team partnered with the Wikimedia Foundation to develop solutions for reducing personal attacks or toxic comments, in Wikimedia editing (Wulczyn, Taraborelli, Thain, & Dixon, 2017). Nobata outperformed state-of-the-art deep learning approaches with their supervised learning approach using a combination of linguistic, n-gram (including character n-grams), syntactic (POS), and semantic (using comment embeddings similar to word2vec) features. In addition, the Yahoo team has released the longitudinal New Feed data set used in the study on (Webscope, n.d.). Wulczyn demonstrated that their machine models can perform as well as three human graders in identifying toxic comments in Wikipedia editing wars, and in addition released the Perspective API to enable developers to utilize their solution. However, see (Hosseini, Kannan, Zhang, & Poovendran, 2017) for comments on adversarial attacks and the resultant fragility of the model - and other models that depend on token-level features. We extend these results and others by developing a system architecture for crowdsourcing sample labeling, a crosssocial-media-platform dataset, and providing an open source classifier for developers to build upon. The classifier is intended to be open sourced in Summer 2017.

3 Methods

In this work, we build upon our initial results using version 1.0 of our dataset see (reference: Harassment detection: a benchmark on the #HackHarassment dataset (Bastidas, Dixon, Loo, & Ryan, 2016). We followed a supervised classification method that uses a data with gold-standard labeled comments and a set discriminating linguistic properties, or features, of each comment to predict the class membership of new or untrained comments. Our features consisted primarily of n-gram and a small set of linguistic features on datasets drawn from The Guardian, Reddit, and Twitter. We performed no significant pre-processing on the data other than tokenization, though in the future we anticipate adding further feature-reduction steps, such as stemming, to improve model performance.

4 Data Source Selection

Three initial data sources were selected: The Guardian, Reddit, and Twitter. Text from each data source were extracted in several ways in Summer 2016. Comments on polarizing or hot-button news articles were extracted from the Guardian, an online news source. Comments from Reddit, a popular social media site, were selected from comment which had received at least 100 down votes. Short texts from Twitter, tweets were hand-curated from an initially machine-selected data set from Twitter, and then further tweets scraped by searching on polarizing or hot-button topics.

4.1 Reddit

Comments from Reddit were downloaded from a publicly available dataset on Google BigQuery, reddit_comments_all_2015. These comments were then filtered to those that had received at least 100 down votes. We used our initial version of the classifier to label these comments. The resulting 5700 harassing comments were then further manually labeled by an in-house team of analysts. Analysts were given instructions and examples for annotation of harassment or non-harassment. In addition, the raters were provided with an additional set of more fine-grained labels but instruction on annotation was not provided.). Each post was labeled independently by at least five Intel Security Web Analysts. A perfect consensus was relatively rare, and so we rated a post as harassing if 40%, 2 of our 5 raters, consider it to be harassing.

4.2 Twitter

Data were comprised of two sources: manual curation and annotation of a pre-existing machine-annotated dataset and a set of scraped tweets using proprietary sampling methods. The sampling

should not be considered unbiased. The initial 5000 tweets were sourced from an online repository of tweets at. Additional tweets were scraped directly from Twitter during July 2016 using a custom twitterbot that queried on hot-button topics as keywords to the Twitter API. These additional tweets were first labeled by our early classifier and then manually labeled by our team (Hart 2016).

4.3 The Guardian

Comments were scraped from 15 articles covering hot-button or polarizing topics. We believe that minimal harassing comments were found in the Guardian dataset as Guardian comments are curated by a team of moderators in accordance with their content policy. Therefore, minimal or no harassing comments should be expected, as we confirmed in the dataset.

Figure 1 shows that the current data set is reason-ably unbalanced overall with a 1:4 ratio of non-harassing to harassing comments. In addition, the categories are unbalanced across source as well as category within source, such that Reddit, de-spite being only 28% of the total comments con-tributed 56% of the harassing comments.

As shown in Figure 2 and Figure 3, average agreement is below 90% for the Guardian and Twitter surveys, with an average across all Qualtrics surveys only .875. This is well below what is typically suggested for raw agreement scores.

Guardian URLs
https://www.theguardian.com/discussion/p/4pcq2
https://www.theguardian.com/discussion/p/4pgek
https://www.theguardian.com/discussion/p/4an9q
https://www.theguardian.com/discussion/p/4p76x
https://www.theguardian.com/discussion/p/4pdqd
https://www.theguardian.com/discussion/p/4phck
https://www.theguardian.com/discussion/p/4pf70
https://www.theguardian.com/discussion/p/4pfe3
https://www.theguardian.com/discussion/p/4k4tx
https://www.theguardian.com/discussion/p/4pd76
https://www.theguardian.com/discussion/p/4jmg2
https://www.theguardian.com/discussion/p/4pg57
https://www.theguardian.com/discussion/p/4p6dt
https://www.theguardian.com/discussion/p/4p6gn
https://www.theguardian.com/discussion/p/4pgbx

Table 1: Guardian URLs used to scrape initial comments.

5 Data Ingest and Annotation Methods

Data ingest process and annotation were heterogeneous in nature. Manual curation was combined with machine annotation in several iterated steps to produce a final annotated dataset. The comment dataset was simply annotated with a Boolean indicating harassment. Harassment was determined on the gold data through a percent voting method: the reported metrics are for 40% and above simple agreement among raters that a given comment is harassment.

All preprocessing, training and evaluation was carried out in Python, using the popular SciKitLearn (for feature engineering and linear models) in combination with Numpy3 (for matrix operations) (Pedregosa et al., 2011; van der Walt, Colbert, & Varoquaux, n.d.).

6 Feature Selection

Features were generated by tokenizing each comment, hashing the resulting n-grams, and computing a TF/IDF value for each token. The resultant feature vectors were used to train a Random Forest classifier. We used the following features:

- Unigram and Bigram TF-IDF: this is a standard feature used in text-categorization. We used unigrams and bigrams. Trigrams were not used because the size of the dataset meant almost all trigrams were too rare for their presence and absence to reach statistical significance.

- Character N-Gram TF-IDF from 3 to 6 characters: The goal with this was to target common alternative spellings of words, particularly frequent in online communication.

- Unigram Token Count: we utilized NLTKs Twitter Tokenizer to tokenize the tokens and count the number of tokens. The Twitter Tokenizer handles URLs and Hashtags much better than a standard punctuation based tokenizers found in NLTK or Sck-kit Learn. Our assumption behind using token count is that harassing texts tend to be brief assaults rather than long diatribes.

- Source: In combination with the token count, we selected a dummy coefficient (toggled as 1 or 0) to highlight if a comment is sourced from Twitter or not.

- Sentiment Polarities: we utilized NLTKs VADER Sentiment Analyzer to generate sentiment polarities for positive, neutral, and negative sentiment. Our assumption was that harassing comments tend to have more negative sentiment, whereas non-harassing comments tend to have more positive sentiment.

7 Training Dataset

The current training dataset contains: 20,432 unique comments. Of these comments, 4136 are labeled as harassment, 16296 are labeled as non-harassment. 12,049 comments are sourced from Twitter, with the remaining 8383 being from Reddit or the Guardian.

8 Machine Learning Model

Bastidas tested a variety of algorithms, including SVM, Decision Tree, Random Forest (ensemble Decision Trees), and Multinomial Nave Bayes (Bastidas et al., 2016). We increased the size of the Reddit dataset and included labeled comments that were sampled from Twitter and The Guardian. We collected performance results using this larger, cross-platform dataset, described in the Data Source Selection section, and a Scikit-Learn Random Forest classifier. For our hyperparameters we limited the number of trees to 200 and left the tree depth unbounded. Subsequently, the data were trained on the Random Forest by splitting the dataset into 80/20 training and evaluation sets, and then the training data were further split into Kfold (n=10) folds for cross-validation and the average results reported in Table 2. Of primary concern to us is to optimize for high recall. We want to minimize our false-negative rate for harassment.

Class	Precision	Recall	F1 Score
Not Harassing	0.93	0.95	0.94
Harassing	0.75	0.68	0.71
Average	0.89	0.90	0.90

Table 2: Random forest classifier results.

9 Future Work

New work from Facebook and OpenAI on text classification suggests obvious next steps. Bytelevel deep neural nets are capable of state-of-theart results on large datasets, can exploit unlabeled data, as described in recent work from OpenAI (Radford, Jozefowicz, & Sutskever, 2017) and have the potential to resist the "adversarial" tokens described in (Hosseini, Kannan, Zhang, & Poovendran, 2017). Using OpenAI's approach with a large, unlabeled dataset for pre-training is an obvious next step. A contrasting approach that requires further evaluation is the FastText model from Facebook's Advanced Research Lab, which, as described in (Joulin, Grave, Bojanowski, and Mikolov, 2016) and (Bojanowski, Grave, Joulin, and Mikolov, 2016), is competitive with deep convolutional neural networks and can exploit unlabeled data using pre-trained WordVectors, while requiring vastly less training time than competitive alternatives.

10 Conclusion

We have presented to our cross-platform harassment dataset, machine learning model. We intend to open our labeling platform to the public to expand the Hack Harassment cross platform dataset. As we complete further rounds of labelling with a public audience, later iterations of this dataset will increase the available samples by at least an order of magnitude, enabling corresponding improvements in the quality of machine learning models for harassment detection. We look forward to both the availability of a larger, cross-socialmediaplatform harassment dataset and seeing the development of classifiers that improve upon our work. We welcome partners able to contribute to expanding the dataset and improving the modeling.

10.1 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from LaTeX using the *pdflatex* command. If your version of LaTeX produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one**

where it was created. Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying "Output to a file", then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. When working with `dvips`, for instance, one should specify `-t a4`. Or using the command `\special{papersize=210mm,297mm}` in the latex preamble (directly below the `\usepackage` commands). Then using `dvipdf` and/or `pdflatex` which would make it easier for some.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs as soon as possible.

10.2 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:
Papers should not be submitted on any other paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

10.3 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

10.4 The First Page

Center the title, author's name(s) and affiliation(s) across both columns. Do not use footnotes for affiliations. Do not include the paper ID number assigned during the submission process. Use the two-column format only when you begin the abstract.

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	11 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 3: Font guide.

Title: Place the title centered at the top of the first page, in a 15-point bold font. (For a complete guide to font sizes and styles, see Table 3) Long titles should be typed on two lines without a blank line intervening. Approximately, put the title at 2.5 cm from the top of the page, followed by a blank line, then the author's names(s), and the affiliation on the following line. Do not use only initials for given names (middle initials are allowed). Do not format surnames in all capitals (*e.g.*, use "Mitchell" not "MITCHELL"). Do not format title and section headings in all capitals as well except for proper names (such as "BLEU") that are conventionally in all capitals. The affiliation should contain the author's complete address, and if possible, an electronic mail address. Start the body of the first page 7.5 cm from the top of the page.

The title, author names and addresses should be completely identical to those entered to the electronical paper submission website in order to maintain the consistency of author information among all publications of the conference. If they are different, the publication chairs may resolve the difference without consulting with you; so it is in your own interest to double-check that the information is consistent.

Abstract: Type the abstract at the beginning of the first column. The width of the abstract text should be smaller than the width of the columns for the text in the body of the paper by about 0.6 cm on each side. Center the word **Abstract** in a 12 point bold font above the body of the abstract. The abstract should be a concise summary of the general thesis and conclusions of the paper. It should be no longer than 200 words. The abstract text should be in 10 point font.

Command	Output	Command	Output
{\"a}	ä	{\c c}	ç
{\^e}	ê	{\u g}	ğ
{\'i}	ì	{\l}	ł
{\.I}	İ	{\~n}	ñ
{\o}	ø	{\H o}	ő
{\'u}	ú	{\v r}	ř
{\aa}	å	{\ss}	ß

Table 4: Example commands for accented characters, to be used in, *e.g.*, BIBTEX names.

Text: Begin typing the main body of the text immediately after the abstract, observing the two-column format as shown in the present document. Do not include page numbers.

Indent: When starting a new paragraph. Use 11 points for text and subsection headings, 12 points for section headings and 15 points for the title.

10.5 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections.

Citations: Citations within the text appear in parentheses as (Gusfield, 1997) or, if the author's name appears in the text itself, as Gusfield (1997). Using the provided LATEX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (Aho and Ullman, 1972), but write as in (Chandra et al., 1981) when more than two authors are involved. Collapse multiple citations as in (Gusfield, 1997; Aho and Ullman, 1972). Also refrain from using full citations as sentence constituents.

We suggest that instead of

> "(Gusfield, 1997) showed that ..."

you use

> "Gusfield (1997) showed that ..."

If you are using the provided LATEX and BibTEX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

If the BibTEX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LATEX package. To disable the hyperref package, load the style file with the `nohyperref` option: `\usepackage[nohyperref]{acl2017}`

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. As of 2017, we are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use BibTEXrecords that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at `http://aclanthology.info/`.

As examples, we cite (Goodman et al., 2016) to show you how papers with a DOI will appear in the bibliography. We cite (Harper, 2014) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, *e.g.*,

> "We previously showed (Gusfield, 1997) ..."

should be avoided. Instead, use citations such as

> "Gusfield (1997) previously showed ... "

Please do not use anonymous citations and do not include acknowledgements when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices, unless they contain references. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983). Use of full names for au-

output	natbib	previous ACL style files
(Gusfield, 1997)	\citep	\cite
Gusfield (1997)	\citet	\newcite
(1997)	\citeyearpar	\shortcite

Table 5: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

thors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (for Computing Machinery, 1983).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

Appendices: Appendices, if any, directly follow the text and the references (but see above). Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

10.6 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 points text. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

10.7 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions of the figures and tables below the body, using 11 point text.

10.8 Accessibility

In an effort to accommodate the color-blind (as well as those printing to paper), grayscale readability for all accepted papers will be encouraged. Color is not forbidden, but authors should ensure that tables and figures do not rely solely on color to convey critical distinctions. Here we give a simple criterion on your colored figures, if your paper has to be printed in black and white, then you must assure that every curves or points in your figures can be still clearly distinguished.

11 Translation of non-English Terms

It is also advised to supplement non-English characters and terms with appropriate transliterations and/or translations since not all readers understand all such characters and terms. Inline transliteration or translation can be represented in the order of: original-form transliteration "translation".

12 Length of Submission

The ACL 2017 main conference accepts submissions of long papers and short papers. Long papers may consist of up to eight (8) pages of content plus unlimited pages for references. Upon acceptance, final versions of long papers will be given one additional page – up to nine (9) pages of content plus unlimited pages for references – so that reviewers' comments can be taken into account. Short papers may consist of up to four (4) pages of content, plus unlimited pages for references. Upon acceptance, short papers will be given five (5) pages in the proceedings and unlimited pages for references.

For both long and short papers, all illustrations and tables that are part of the main text must be accommodated within these page limits, observing the formatting instructions given in the present document. Supplementary material in the form of appendices does not count towards the page limit.

However, note that supplementary material should be supplementary (rather than central) to the paper, and that reviewers may ignore supplementary material when reviewing the paper (see Appendix **??**). Papers that do not conform to the specified length and formatting requirements are subject to be rejected without review.

Workshop chairs may have different rules for allowed length and whether supplemental material is

[1] This is how a footnote should appear.

[2] Note the line separating the footnotes from the text.

welcome. As always, the respective call for papers is the authoritative source.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery* 28(1):114–133. https://doi.org/10.1145/322234.322243.

Association for Computing Machinery. 1983. *Computing Reviews* 24(11):503–512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/P16-1001.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. http://aclweb.org/anthology/C14-1001.