The Online Harms Observatory Methodology for Detecting Abuse

The Alan Turing Institute Public Policy Programme, Online Safety Team

January 2022

Offensive content warning: This note contains some examples of abuse (all are synthetic, i.e. not real). We have sought to minimize the number of examples which are included but some are needed for our research purposes. Please be aware that you might find them offensive and they could cause you distress.

This Note outlines the methodology for training Artificial Intelligence (AI) to automatically detect online abuse, developed by The Alan Turing Institute Public Policy Programme's Online Safety Team and used to power the Online Harms Observatory (OHO). The first section summarises the background of the OHO and provides an introduction to using machine learning for detecting abusive language. We then explain how our machine learning models are trained through data-centric techniques.

If you have any questions about this Note or would like more information about either the OHO or how we detect online abuse, please reach out to Dr. Bertie Vidgen, bvidgen@turing.ac.uk.

Background

About

The OHO is a new analytics platform from the Alan Turing Institute's Public Policy Programme. It combines large-scale data analysis and cutting-edge AI developed at The Turing to provide real-time insight into the scope, prevalence and dynamics of harmful content online. It aims to help policymakers, regulators, security services and civil society stakeholders better understand the landscape of online harms. Initially, it will focus on online hate, personal attacks, extremism and misinformation. The OHO is supported by DCMS.

The OHO comprises several trackers, each of which collects data on a specific set of online actors or community. The trackers stream data in real-time from social media platforms, which is analysed using a range of statistical and machine learning tools from The Turing. The results are then visualised, providing accessible and interactive insights into the prevalence, trends, spread, and dynamics of online harms. Initially, we are collecting data from Twitter and will then expand to Reddit, and a range of niche platforms. The first two trackers are: (a) Premier league footballers and (b) MPs.

As of January 2022, we have completed a demo version of the OHO (v0.1). We have engaged with stakeholders, developed our backend infrastructure, and launched data collection. We are now working to build v1 of the OHO in time for AI UK on March 22/23 2022. By this point we aim to use our own AI, as described in this Note, in the OHO.[1]

Detecting abuse with machine learning

Finding and categorising abuse is a difficult task – humans routinely disagree about how the term 'abuse' should be defined; where the line should be drawn between abusive and non-abusive content to protect free expression; and how complex issues like intention, irony, humour and context should be handled.

Unsurprisingly, detecting abuse with AI is a technically difficult task and even state-of-the-art models can have considerable weaknesses.[2] Models routinely miss very obvious forms of hate and abuse, fail to pick up on coded language and slangs, and flag legitimate, non-abusive speech as hateful. Both these types of errors can have serious consequences in the real-world, either leaving people at risk of serious harm or limiting free expression. Research also shows that AI systems for abuse detection can be biased, brittle, unexplainable, and can quickly go out of date and/or be easily tricked by adversarial attacks.

Despite its challenges, machine learning systems have key advantages over human-review: it is scalable, cost-efficient, improvable and consistent. In contrast, having humans review entries is time-intensive and expensive. It does not scale because handling more content generally requires a linear expansion in the number of reviewers; this is particularly problematic when more capacity is required at very short notice (such as in the aftermath of a terror attack or political event). Human review is also inconsistent, with the same reviewers giving different assessments at different points of time – they also face a serious risk of harm, potentially suffering from emotional and mental health problems from viewing abusive content.

Simplistic computational tools for detecting abuse, such as searching for abusive key terms and phrases, are scalable but suffer from different but equally substantial problems. They cannot take into account context, sentence structure and nuance, and are inflexible to new forms of abuse.

Types of abuse tracked by the OHO

Defining abuse is a challenging social task, with considerable disagreement amongst experts. In the OHO we are training machine learning classifiers to automatically detect two types of abuse in social media content[3]:

1. Person-directed abuse: Content which directs intense negativity against an identifiable person in the tweet. It includes serious character based attacks, such as accusing the person of lying, as well as aggression, insults and menacing language. Examples include:
    a. "Fuck off @user"
    b. "Trump is a massive bellend"

2. Identity-directed abuse: Content which directs negativity against an identity (i.e. 'hate speech'). An 'identity' is a social category that relates to a fundamental aspect of individuals' community, socio-demographics, position or self-representation (including but not limited to race, religion, gender and sexuality). Identity-directed abuse

[1] To find out more and purchase tickets go to: https://www.turing.ac.uk/ai-uk

[2] See our prior research for a more detailed discussion of the challenges of detecting online abuse: https://aclanthology.org/2021.acl-long.132/ https://aclanthology.org/2021.acl-long.4/ https://aclanthology.org/W19-3509/

[3] This draws on conceptual and machine learning analyses in our prior research: https://aclanthology.org/2021.naacl-main.182/

includes derogatory, threatening, inciting, insulting and demeaning remarks. Examples include:

a. "[GROUP] cant speak English, they're savages"
b. "[GROUP] are trash, they make me sick"

Methodology

There are three main aspects to training AI. First, is data acquisition and how we identify and label relevant examples. A large dataset which lacks diversity or has low-quality labels is often not very useful for AI. Second, is the modelling and how we actually train the AI. Third, is evaluation and how we ensure that – before models are deployed 'live' – they perform to the required standard, and biases and limitations have been both documented and mitigated.

Data acquisition

Data acquisition refers to the process of sampling, labelling and using data to produce machine learning models. Data is crucial to how we train our machine learning models – we have a workstream called *Data-centric AI* precisely because, whilst the models are themselves important, for complex social tasks like abusive content detection the data is often the most important aspect: Garbage leads to Garbage Out.[4] We use an iterative approach to modelling which is a far more efficient way of collecting data to address model weaknesses. These are also known as 'human-and-model in the loop' approaches for acquiring data, where we keep retraining and improving our models, using the models' performance as a signal to direct the new data acquisition.

We primarily use two human-and-model in the loop techniques for identifying new data (and training models).

1. Active learning: A simple model is trained on a corpus of existing data (both in-house datasets and open source datasets). It is then applied to a large sample of relevant data (e.g. tweets from one of the trackers), and we use statistical models to identify entries which 'confuse' the model. These are then sent for annotation. The model is retrained on the annotated data and the process is repeated. *In effect, the model chooses itself what data it needs to learn from.*

2. Adversarial learning: Similarly to active learning, we start by training a simple model on a corpus of existing data. We then task annotators with creating synthetic entries which they think will trick the model. For instance, many hate speech models overfit to identity referents, and can be tricked by statements such as "I hate the black tiles in my kitchen". The model is retrained on the annotated data and the process is repeated. This method is very effective as annotators find and exploit the model's key limitations. However, the challenge is that the entries are all synthetic, which can introduce some biases. *In effect, the model learns from what it is worst at classifying.*

Effective labelling of data that has been sampled is essential. If entries with similar semantic expression (i.e. two tweets which use the "N word" in a very similar way) are labelled differently than models learn from contradictory signals which makes it hard to

[4] See: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0243300

identify an appropriate decision boundary.[5] We typically adopt an expert-led approach, in which annotation guidelines are carefully constructed by subject matter experts, and then every entry is labelled three times by a large team of annotators. In cases where the three annotators disagree on the correct label, we send the results to a more senior annotator to make a final decision or collect more annotations (in some cases, as many as 9 annotations). We follow best practice guidelines for ensuring annotator wellbeing and safety, including providing access to support services, where needed.[6]

Modelling

We use state-of-the-art transformer models, which are a form of deep learning. These are large language models which have been pre-trained on large online corpora, and have been shown to perform exceptionally well at abuse detection.[7] The creators of the models (typically large tech firms such as Google, Microsoft and Facebook) train them using a process called 'self-supervision' on amounts of online user-generated data, often at a cost of millions or tens of millions of dollars. The models are very complex, and can have billions, or even trillions, of parameters.[8] Most large language models are open sourced, which means that we can use them in downstream applications, such as abuse detection. To use the models for this task we 'fine tune' them on a newly labelled domain specific dataset. This is where the top layers of the model are retrained on the new data, quickly adapting to the specific task we want them to perform at (in our case, detecting abusive language). Performance is often ~10 percentage points higher when using the transformer based approach, compared against machine learning models trained from scratch.

Evaluation

Evaluating abusive content detection models is difficult because the task is itself highly subjective and multi-faceted.[9] In a traditional machine learning approach models are evaluated by taking a portion of the training dataset, typically around 20%, and keeping it 'held-out', i.e. not letting the model see it in training. This held-out portion is called the 'test set'. Aggregate statistics can then be calculated. Within this traditional approach, our machine learning models are assessed against two main criteria, Precision and Recall. We generally aim to maximise precision first because this means that our results can be trusted. We then improve the model's coverage to maximise recall.

1. Precision: Of all the content which the model identifies as abusive, how much actually is abusive? For instance, if the model identifies 100 posts as abusive, what percentage of them are "true positives", i.e. are genuinely abusive.

2. Recall: Of all the content which is abusive, how much does the model identify? For instance, if there are 300 abusive posts in our dataset, what percentage of them will be correctly identified by the model?

The traditional approach to model evaluation has increasingly been criticised for not

[5] For a discussion of decision boundaries in machine learning see: https://aclanthology.org/2020.findings-emnlp.117/

[6] See a set of guidelines for annotator wellbeing that we released publicly: https://github.com/bvidgen/Challenges-and-frontiers-in-abusive-content-detection

[7] See an overview to transformer-based machine learning via BERT models: https://arxiv.org/pdf/2002.12327.pdf

[8] See: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

[9] See the work of Dynaboard, which proposes model evaluation in terms of performance, robustness, fairness and scalability: https://arxiv.org/abs/2106.06052

providing fine-grained insight into the strengths and weaknesses of models, limiting our ability to improve them. We address this constraint in three main ways:

1. Secondary labels: Our datasets are annotated with additional labels. For instance, we record both the target of hate (i.e. which group has been attacked) and the type of hate (i.e. whether it is derogatory, threatening, etc.). This means that we can evaluate model performance against specific types and targets, identifying any limitations or biases in how they perform.

2. Error analyses: We conduct qualitative ground-up analyses of the held-out test set, identifying consistent errors that our models are producing.[10]

3. Functional testing: This is a new approach to machine learning model evaluation, inspired by software product testing. We start by identifying clear functionalities that we want the model to have, such as correctly identifying threatening language and hate expressed via leet speak, as well as not identifying negativity against inanimate objects as hate (e.g. "I hate this table"). We then construct clear-cut examples which correspond to each functionality and test the model against them.[11]

Abuse detection in the OHO demo (v0.1)

We have developed a static demo of the OHO to help solicit stakeholder feedback (v0.1). In the demo, we have deployed the commercial 'toxicity' model from Google/Jigsaw's Perspective API.[12] It is widely used for academic research and content moderation, announcing at the start of 2021 that it receives over 500 million calls every day.[13]

We aim to release v1 of the OHO by the end of Q1 2022. We will have custom-trained models in place to detect both Identity-directed and Person-directed abuse.

[10] See the final parts of our paper on detecting East Asian prejudice, where we created a new dataset and trained a set of machine learning models: https://aclanthology.org/2020.alw-1.19/

[11] See two papers that we have produced for functional testing of hate speech models: HateCheck and HatemojiCheck https://aclanthology.org/2021.acl-long.4/ and https://arxiv.org/abs/2108.05921

[12] See: https://www.perspectiveapi.com/

[13] See: https://www.prnewswire.com/news-releases/googles-jigsaw-announces-toxicity-reducing-api-perspective-is-processing-500m-requests-daily-301223600.html