Unintended Bias in Toxicity Classification

Abdul Rahman Gulam Mohammed Hussain
Briana Haldaman
Maryam Hashemitaheri
Satya Kotha
Shilpa Patil

Introduction

- In 2019, there are 3.2 billion social media users, roughly 42% of the world population.
- People have a greater ability to express themselves freely with little consequence
- Toxicity detection can help to identify and eliminate users who negatively impact others
- When AI teams began building models to detect toxicity, bias was inadvertently built in because it already existed in the training datasets

Toxicity is defined as anything *rude, disrespectful or otherwise likely to make someone leave a discussion*.

Problem Statement

1 in 3 young people experience cyberbullying

Cyber-bullying takes many forms, some of which include sending mean messages, spreading rumors, posting harmful or threatening messages and photos.

Victims can suffer from anxiety and depression into their adult years

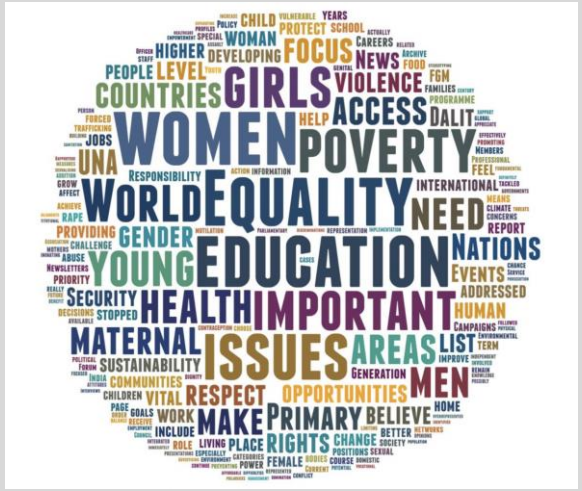
The cyberbullies tend to engage in additional behaviors such as alcohol and drug abuse, early sexual activity, and abuse towards their significant other.

Bias in the AI models used by companies using them to drive business decisions, Risk analysis, outcome predictions

Business Goal

- To develop a strategy to reduce unintended bias in machine learning models and build models that work well for a wide range of conversations.
- To build a model that recognizes toxicity while minimizing the unintended bias with respect to mentions of identities.

Background



Centre for Artificial Intelligence Research (CAiRE) Hong Kong University of Science and Technology

- Abusive language detection models tend to have problems with bias towards different groups, causing an unbalanced training dataset
- Toxic language detection is important to detect cyber-bullying, hate crimes, and discrimination
- Bias correction is important to improve the robustness of AI models
- As an example: Gender bias cannot be measured when evaluated on the original dataset as the test sets will follow the same biased distribution. A separate dataset must be created for male vs. female

Reducing Gender Bias: https://aclweb.org/anthology/D18-1302

Background

Animesh Koratana & Kevin Hu Stanford University Department of Computer Science

- The rapid growth of online platforms have brought awareness to the frequent use of toxic language
- When someone engages in an online discussion on social media, blogs, or comment sections, they are exposed to the risk of being harassed by other commenters . Instances of offensive comments have negative impacts the dynamics of the online community.
- Toxic language detection is a difficult task that is more than just keyword recognition and pattern identification in the grammar.
- While machine learning has shown significant improvement in this area, it has come at a cost of speed and scalability

Toxic Speech Detection:
http://web.stanford.edu/class/cs224n/reports/custom/15744362.pdf

Data Profile

The training Dataset has are **1804874** rows and **45** variable columns collected from public comments from 2015-2017.
ID = Row ID
Target = Toxicity Label

- Toxicity label is in fraction form and represents the fraction of human raters who believed the attribute applied to the given comment.
- Comments with target value >0.5 are considered to be in positive class for toxicity.

Labeled by annotators

- Annotators were asked to: "Rate the toxicity of this comment"
- Varies from Very Toxic to Not Toxic
- These ratings were then aggregated with the target value representing the fraction of annotations that annotations fell within the former two categories.

Comment_Text = Text from the individual comment

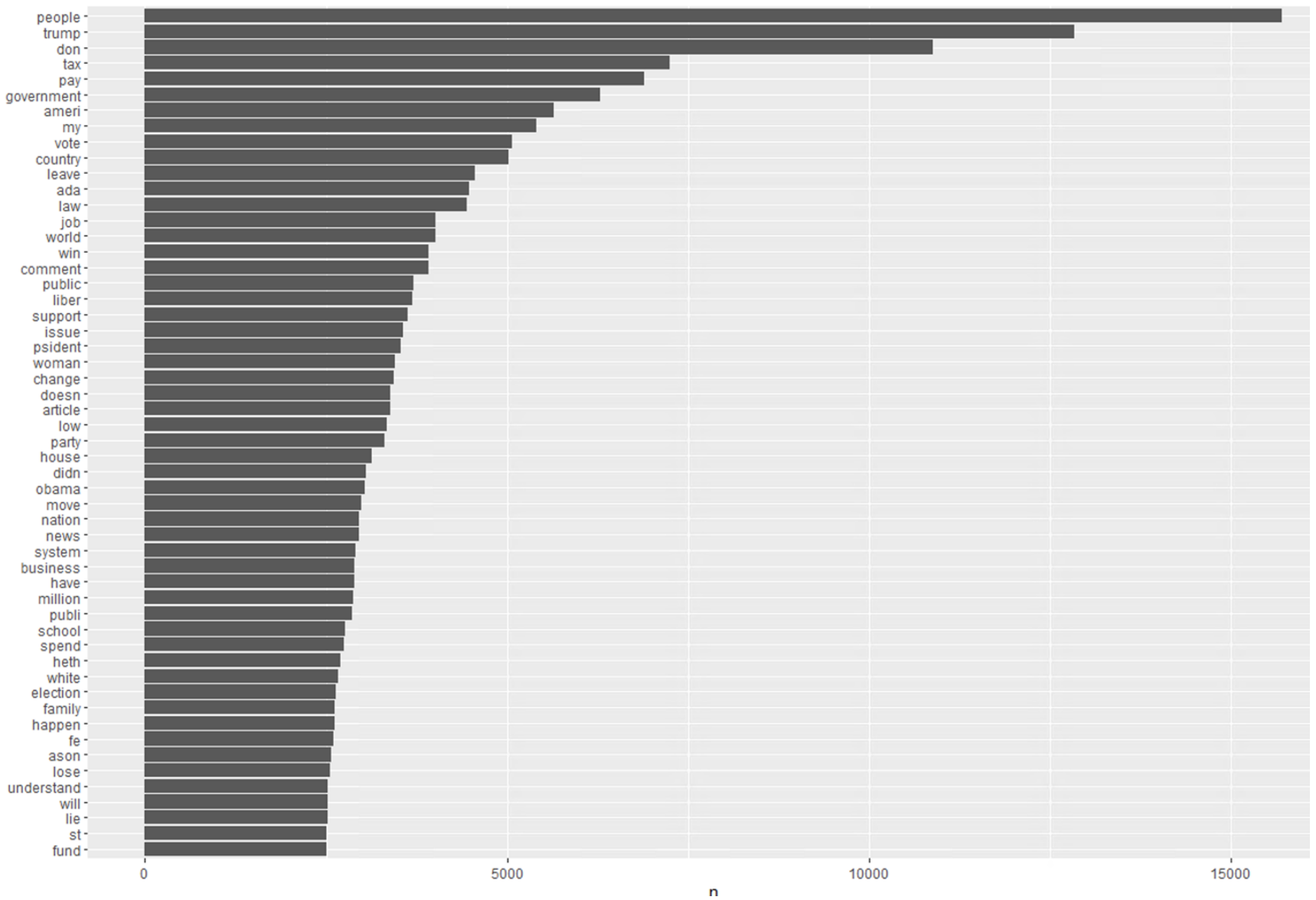
Data Profile

- Data also has identity attributes based on the identity mentioned in the comment.
- Annotators were asked to indicate identities, such as gender, race, sexual orientation, disabilities, etc.
- Those were aggregated to fractional values representing the fraction of raters who said that particular identity was mentioned in the comment.
- Ex. male, female, homosexual_gay_or_lesbian, christian, muslim, jewish, black, white, psychiatric_or_mental_illness
- Only the “id”, “comment_text”, “target” attributes were kept as our test data did not have any of other attributes.

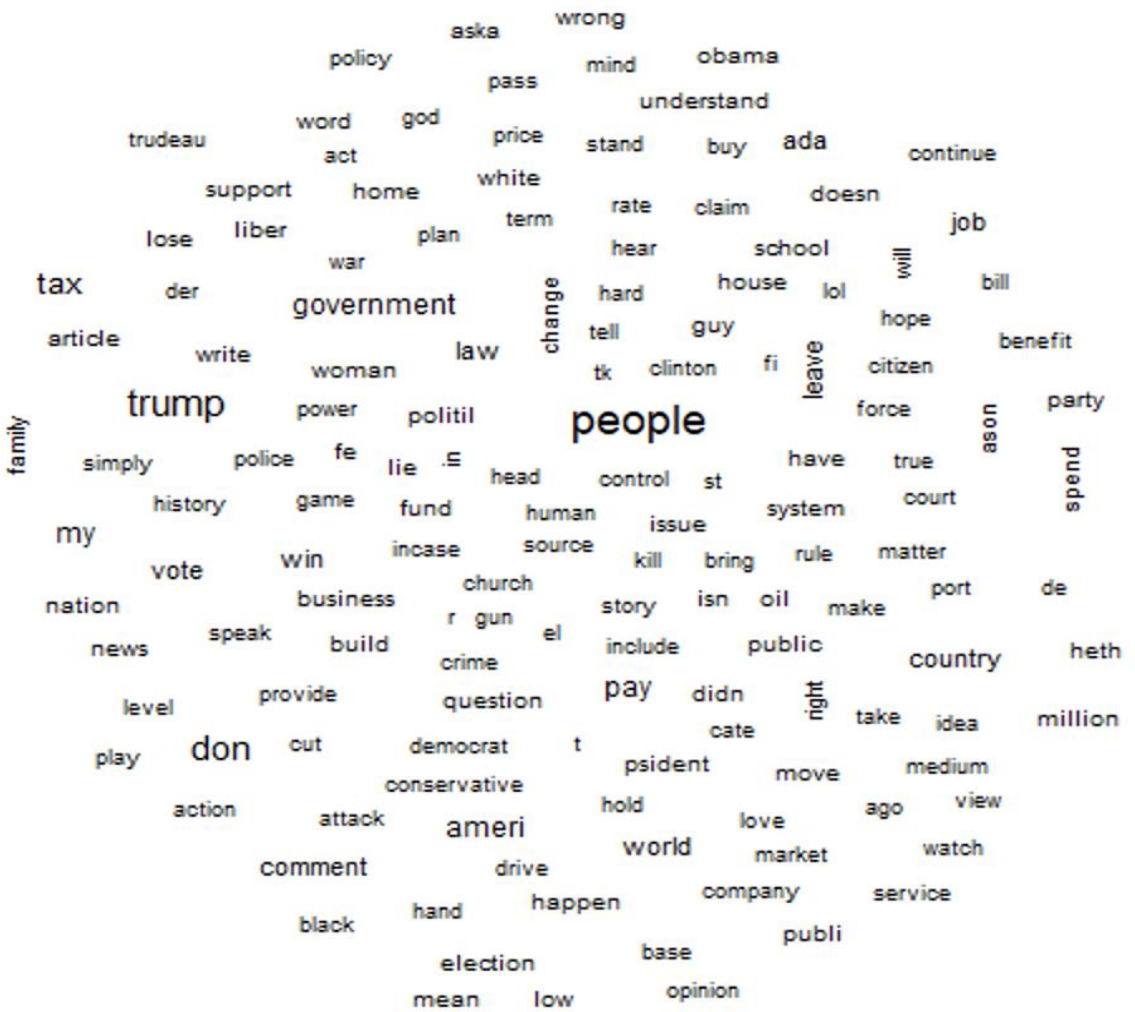
Data Cleaning

- Converted target values > 0.5 to 1 and <= 0.5 to 0
- Comment_text attribute:
  - Removed punctuations, special characters, numbers, stop words
  - converted to lowercase
  - Tokenized
  - Then removed customized list of words (time, my, https, www, http, can, see, etc.)
  - Lemmatized the tokens

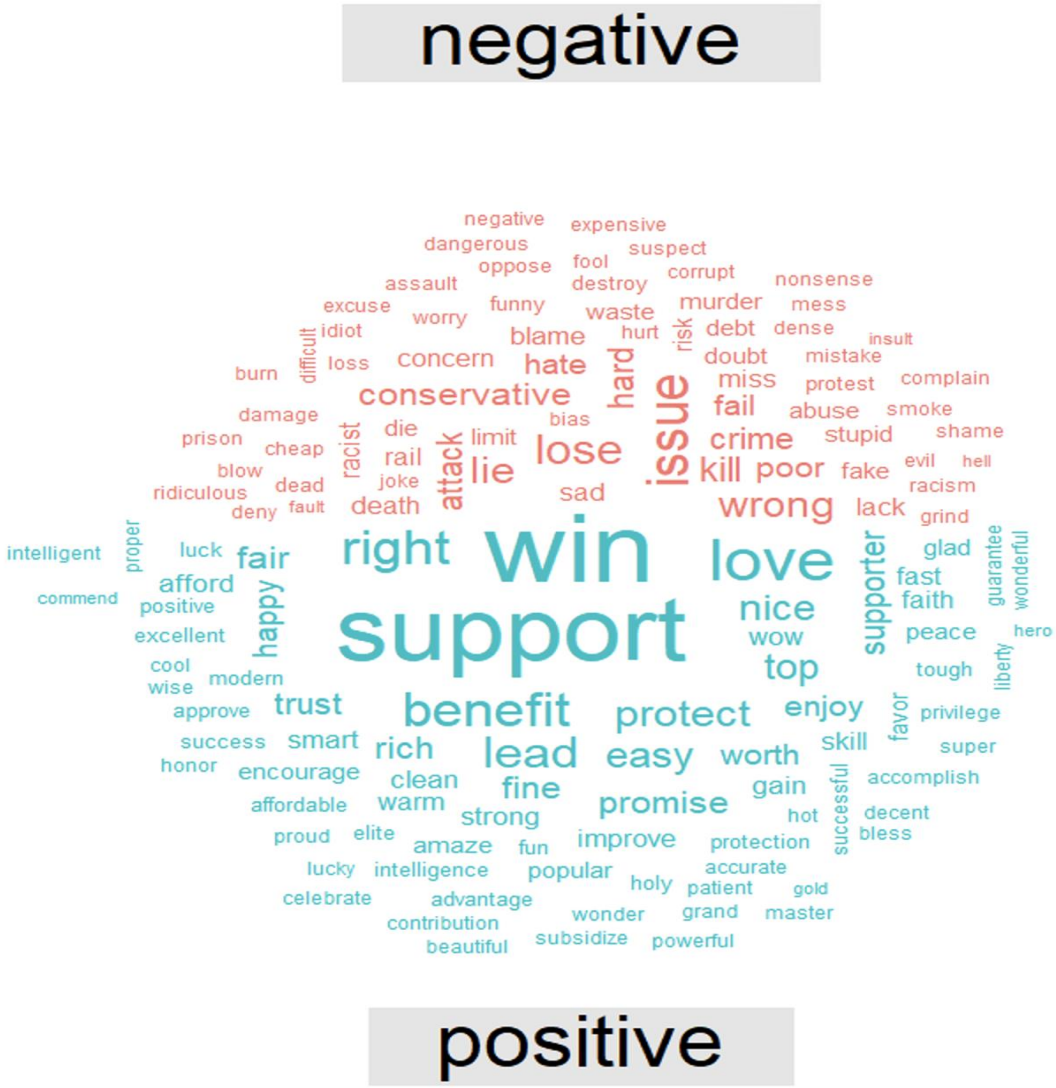
Data Exploration



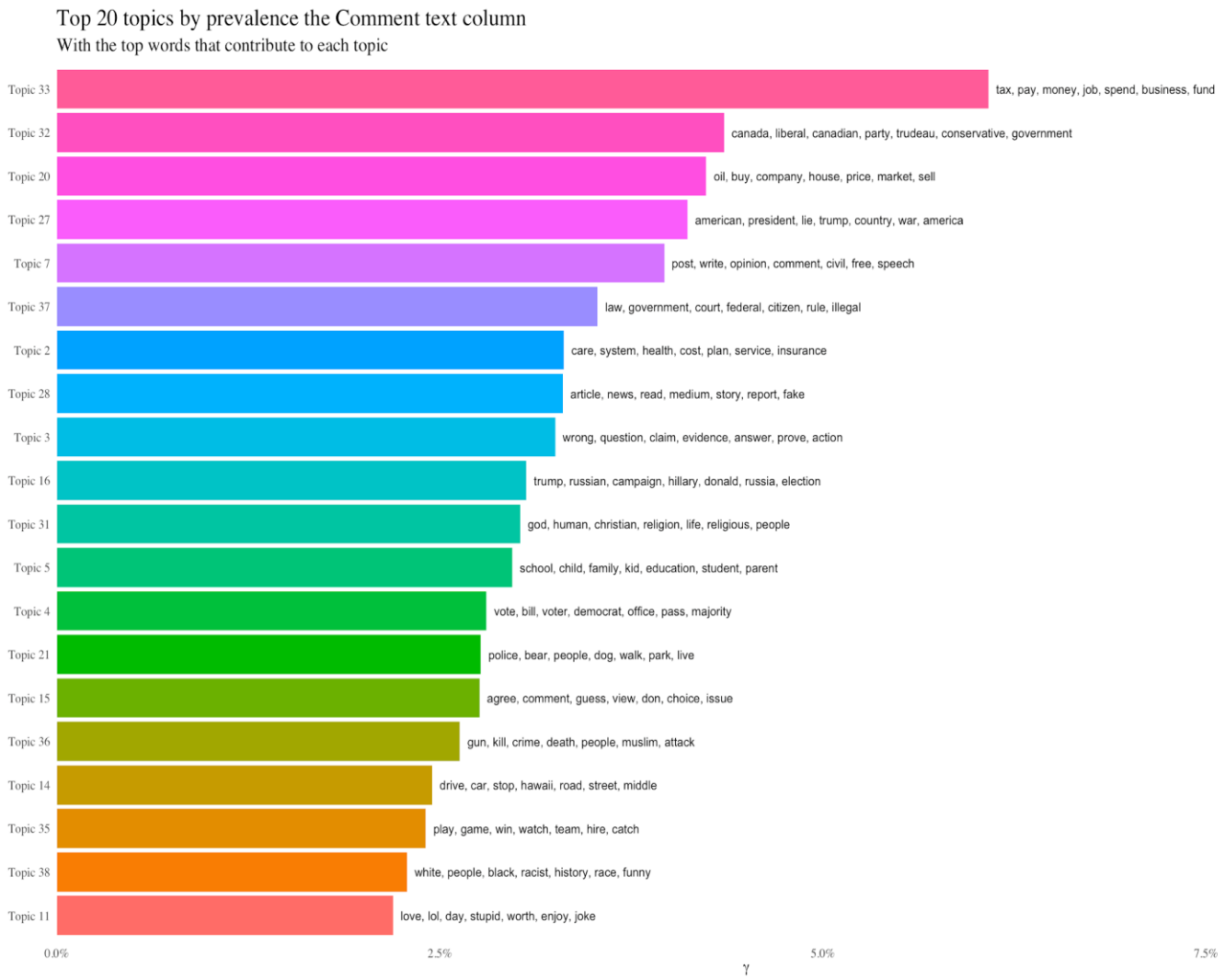
Data Exploration



Data Exploration



Data Exploration



CNN Model

Convolutional Neural Network

Our CNN model contains the following layers:

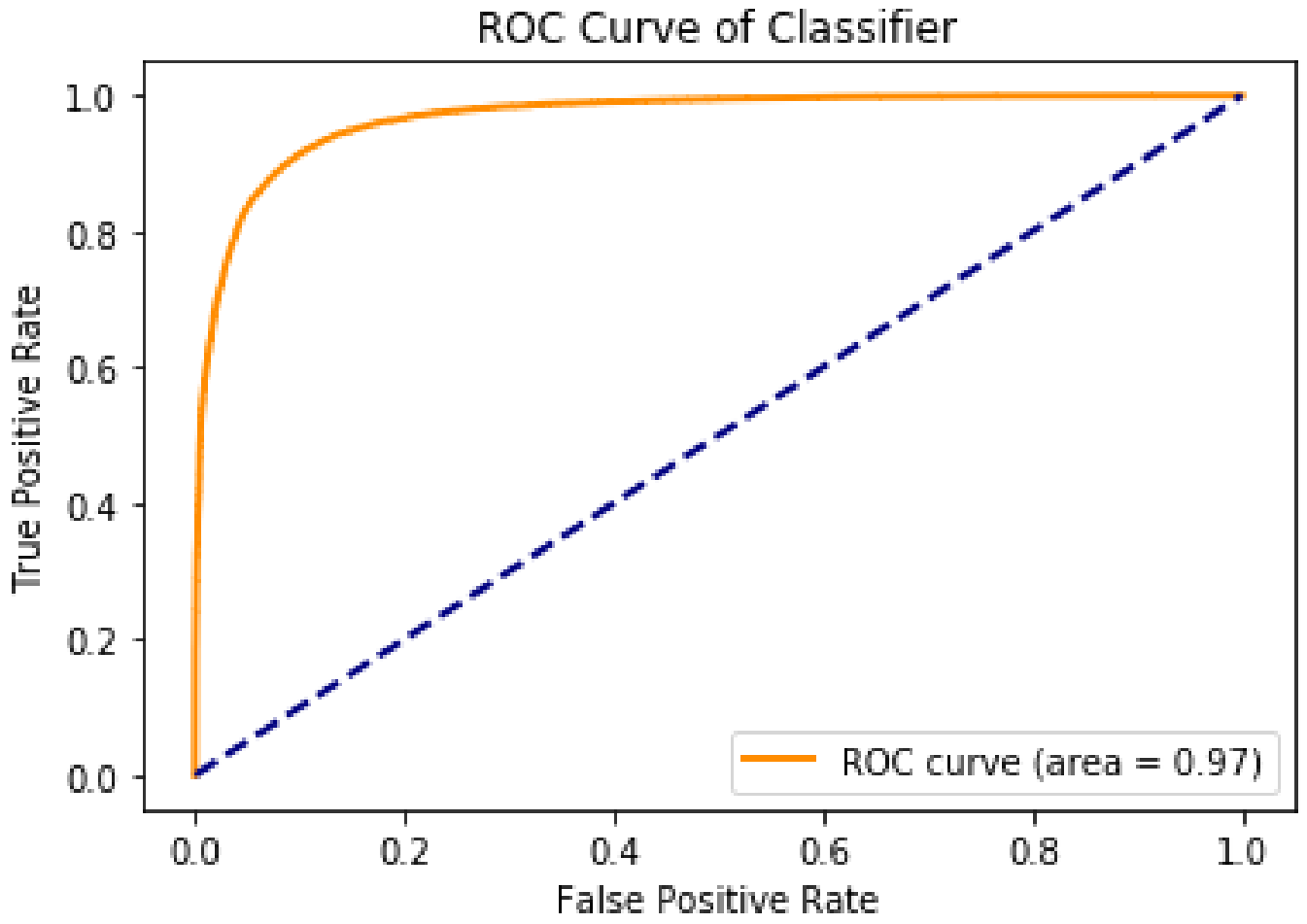
Input ↠Embedding ↠ Convolutional Filters ↠ Drop Out ↠ Dense Layer ↠ Sigmoid ↠ Output

Using pre-trained word embedding model **"ConceptNet NumberBatch",** each record is converted to 300-D numeric vector.

For our CNN model, we use:

- **Pad size = 150**: Only 150 initial words for each record is used
- **Convolutional Filters**: We use filter sizes of 2,3, and 4 (100 filters of each size).
- **Drop out = 40%**: removing 40% of filters output will notably prevent over-fitting.
- **Epochs = 10**: The models iterated 10 times through the data.
- **Batch size = 30**: In each iteration, the model reads the data in blocks of 30 records.

Results



Precision: 0.91
Recall: 0.91
F1 Score: 0.91

Conclusion

- The model can assist social media platforms to minimize the bias in the toxicity of the comments
- The model can identify the real toxic comments and can be removed/eliminated

Questions


WHERE?
WHEN?
WHY?
WHO?
WHAT?
HOW?
WHERE? WHICH? WHOSE? WHEN? WHY?
WHO? WHERE? WHAT? HOW?
WHAT? HOW? WHY?
WHY? HOW?
WHERE? WHAT? HOW?
WHO? WHOSE?
WHERE? WHICH? WHOSE? WHEN? WHY?