Toxicity In Online Comments

David Feier, Evan Sells, Noah Treutel, Evan Kelley, Advisor/Professor: Dr. William Eberle

Introduction

Our goal in this project is to categorize toxic comments used in IRC chat rooms and forum pages via a machine learning model in order to keep communications safe and welcoming to the public. We define toxicity as any message that is rude, disrespectful, or otherwise likely to make someone leave a discussion.

Overview

- Our Data Source(s): Our training data is a set of English data provided by two previous Kaggle Jigsaw competitions, and our test is given to us in Kaggle.
- We initially wanted to classify comments in multiple languages as toxic or not, but due to time constraints and level of knowledge, we had to limit it just English.
- How we are obtaining the data: This data is provided as part of the Kaggle competition we have chosen. Additionally, it is unlikely we will need to gather additional data, as we believe what was supplemented is more than enough.
- Goals of the team's analysis
  - We want to be able to identify a dataset of toxic comments with at least a 70% success rate. Simply put, we want to be able to detect if a comment is toxic (scored as a 1) or non-toxic (scored as a 0) correctly 70% of the time.
  - We want to be able to successfully identify both new and previously existing comments as toxic or non-toxic.

Methods

- Using R and RStudio
  - Classifying Text Using Keras Library
  - Using Tensorflow Library
  - Using dplyr Library
  - Using Tidyverse Library
- Using Anaconda/Miniconda

Training Data

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777bf	Explanation	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really r	0	0	0	0	0	0
0001b41b1c6bb37e	"	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero	0	0	0	0	0	0
00025465d4725e87	"	0	0	0	0	0	0
0002bcb3da6cb337	[redacted] BEFOR	1	1	1	0	1	0
00031b1e95af7921	Your vandalism to th	0	0	0	0	0	0
00037261f536c51d	Sorry if the word 'no	0	0	0	0	0	0
00040093b2687caa	alignment on this su	0	0	0	0	0	0
0005300084f90edc	"	0	0	0	0	0	0
00054a5e18b50dd4	bbq	0	0	0	0	0	0
0005c987bdfc9d4b	Hey... what is it..	1	0	0	0	0	0
0006f16e4e9f292e	Before you start	0	0	0	0	0	0
00070ef96486d6f9	Oh, and the girl abov	0	0	0	0	0	0
00078f8ce7eb276d	"	0	0	0	0	0	0
0007e25b2121310b	Bye!	1	0	0	0	0	0
000897889268bc93	REDIRECT Talk:Voyd	0	0	0	0	0	0
0009801bd85e5806	The Mitsurugi point	0	0	0	0	0	0
0009eaea3325de8c	Don't mean to	0	0	0	0	0	0
000b08c464718505	"	0	0	0	0	0	0
000bfd0867774845	"	0	0	0	0	0	0
000c0dfd995809fa	"	0	0	0	0	0	0
000c6a3f0cd3ba8e	"	0	0	0	0	0	0
000cfee90f50d471	"	0	0	0	0	0	0
000eefc67a2c930f	Radial symmetry	0	0	0	0	0	0
000f35deef84dc4a	There's no need to a	0	0	0	0	0	0
000ffab30195c5e1	Yes, because the mo	0	0	0	0	0	0
0010307a3a50a353	"	0	0	0	0	0	0

Test Data

id	comment_text
00001cee341fdb12	Yo[redacted] Ja Rule is more succesful then you'll ever be whats up with you and hating you sad [redacted]
0000247867823ef7	== From RfC ==
00013b17ad220c46	"
00017563c3f7919a	:if you have a look back at the source, the information I updated was the correct form. I can only
00017695ad8997eb	I don't anonymously edit articles at all.
0001ea8717f6de06	Thank you for understanding. I think very highly of you and would not revert without discussion.
00024115d4cbde0f	Please do not add nonsense to Wikipedia. Such edits are considered vandalism and quickly undo
000247e83dcc1211	:Dear god this site is horrible.
00025358d4737918	"
00026d1092fe71cc	== Double Redirects ==
0002eadc3b301559	I think its crap that the link to roggenbier is to this article. Somebody that knows how to do thing
0002f87b16116a7f	":: Somebody will invariably try to add Religion? Really?? You mean, the way people have
0003806b11932181	, 25 February 2010 (UTC)
0003e1cccfd5a40a	"
00059ace3e3e9a53	"
000634272dd04eb	==Current Position==
000663aff0fffc80	this other one from 1897
000689dd34e20979	== Reason for banning throwing ==
00083476911537 0c	:: Wallamoose was changing the cited material to say things the original source did not say. In res
000844b52dee5f3f	[blocked]] from editing Wikipedia. |
00084da5d4ead7aa	==Indefinitely blocked==
00091c35fa9d0465	== Arabs are committing genocide in Iraq, but no protests in Europe. ==
000968ce11f5ee34	Please stop. If you continue to vandalize Wikipedia, as you did to Homosexuality, you will be blo
0009734200a85047	== Energy ==
00097b6214686db5	:yeah, thanks for reviving the tradition of pissing all over articles because you want to live out yo
0009aef4bd9e1697	MLM Software,NBFC software,Non Banking Financial Company,NBFC software company,NBFC
000a02d807ae0254	@RedSlash, cut it short. If you have sources stating the RoK is sovereign post them. Otherwise pl
000a6c6d4e89b9bc	==========
000bafe2080bba82	.

What Has Been Done

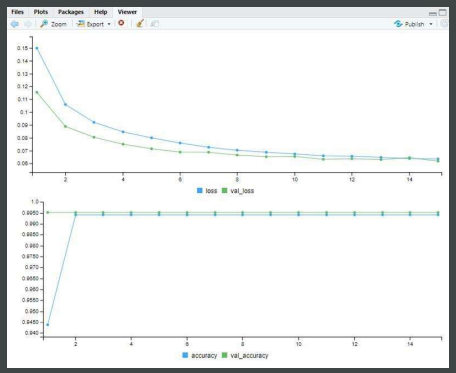
- As of now, we are in the early phases of creating and testing our model. In our first iteration, we are currently seeing signs of overfitting and are looking for ways to change and adapt the model to account for this, all while maintaining high accuracy and lowering loss. Currently, before our model receives the data, we run some text processing to filter out unnecessary characters and homogenize the language to all lower case. From here, we tokenize the data, meaning we take every comment and divide them into individual words in a vector, assigning them unique integer IDs. Following this, we replace the words with these identification numbers that are stored in our word index dictionary.

- Now our data is ready to run through the model. First, the model takes the average of the incoming data and pools them, as this is to account for different comment lengths to prep for the next layers. Next, we pass that data to our ReLU dense layer, which is a piecewise linear function that assigns random weights and categorizes outputs either a positive, negative or no correlation to the data. The last hidden layer in the model is the sigmoid layer, which creates a binary cross-entropy output for our six categorical variables detailing toxicity.

Future Plan

- We plan on looking over our current model to see if our model is overfit and check its validity.
- As we continue to develop improvements in our model, we are looking to reduce our loss past what it is currently at now.
- To help with the accuracy of the data we believe that we would need to create custom weights via word embeddings instead of using ReLU, which assigns random weights.

Current Data Model Results



Summary of Model

```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 250, 6)            108000
_________________________________________________________________
global_average_pooling1d_3 (GlobalAver (None, 6)       0
_________________________________________________________________
dense_7 (Dense)              (None, 16)                112
_________________________________________________________________
dense_6 (Dense)              (None, 6)                 102
=================================================================
Total params: 108,214
Trainable params: 108,214
Non-trainable params: 0
_________________________________________________________________
```

References

https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

https://www.kaggle.com/krosam/why-a-such-low-score-with-r-and-keras

https://www.onceupondata.com/2019/01/21/keras-text-part1/

https://www.onceupondata.com/2019/01/24/keras-text2-multilabel-classification/

Acknowledgement: TTU; Computer Science Department; CSC-3220 Fundamentals of Data Science (Term Project)