Toxic Span Detection

Babafemi G. Sorinolu
School of Systems and Enterprises
Stevens Institute of Technology
Hoboken, USA
bsorinol@stevens.edu

I. Introduction

The use of technology in 21st century generation has made the world a global village. It has created the opportunity for people to have meaningful interactions with each other through forums, chats and comments. However, the ability to interact on the online space, some people tend to post or communicate in hurtful and disrespectful tone to each other as a form of attack, intimidation [3].

To ensure a safe community for everyone, there is need to always censor and ban negative and hurtful comments and messages that are posted online. Due to the large number of people on the internet, it will be impossible for human moderators to control and detect negative words posted on these platforms.

Hence, the usefulness of natural language techniques to help detect negative words in the large text information on the online space.

The potential application of the toxic span detection systems can be used on social media platforms such as facebook, twitter, instgram to ensure a healthy community. It can also be used to block and limit the individuals with racist or threatening behaviors.

Toxic span detection can be a challenging task while implementing due to the following reasons

- Dataset label imbalance: In the dataset, the ratio of the non-toxic words to the toxic words phrases in the sentences is imbalanced. This affects the model performance in detecting the toxic words.
- Dataset size: Training a deep learning model requires a large dataset for effective model performance. To improve the performance of toxic span detection, there will be need obtain more data samples.
- The prediction of toxic word requires contextual meaning.

In this report, I present the approach used in detecting the toxic words in a sentence by using GRU recurrent neural network on a sample dataset containing sentences and their toxic word span.

The structure of this paper is organized as follows: section II describes the task problem formulation in details, section III describes the dataset and method used , section IV presents the results obtained from the experiment and section V concludes the report.

II. Problem Formulation

I consider the toxic span detection problem as a sequence labelling task. Sequence labeling is a type of pattern recognition task that involves the assignment of a categorical label to each word token of a sentence sequence [1].

A popular example of the sequence labeling type of problem is the part of speech tagging. The part of speech tagging assigns a part of speech such as verb, noun, pronoun etc. to each word in a given sentence. In similar manner, we can assign a toxic or non-toxic label to each word in a given sentence and retrieve the word span after the prediction.

Therefore we define the basic formulation as follows:

- **Input:** Sentence e.g "Your comment is ridiculous"
- **Output:** Toxic words , A list of indexes of the toxic words (List[int]) e.g ridiculous , [16, 17, 18, 19, 20, 21, 22, 23, 24, 25]

III. Method

A. Overview of Dataset

The toxic span dataset used in this work was obtained from here. The number of samples contained in the train, test and validation set were 7939, 2000, and 690 records respectively. Fig 1 shows a sample of the two column dataset containing the sentence and a list of indexes of the toxic words in the sentence.

index	spans	text
1985	12,13,14,15,16,17,18,19,20,32,33,34,35	Keep hiring imbeciles like this jerk and you will end up with a no firearms for rent-a-cops bill next session.
1986	0,1,2,3,4,5,6	Asshole.
1987	52,53,54,55,56,57	Calling tens of millions of voters "deplorable" or "stupid" (because they voted for the other guy) is condescending. But hey, feel free. It's working so well for you guys.
1988	161,162,163,164,165,166	"molesting his own daughter and betraying America to Russia" The first part of that sentence makes you a disgusting liar. The second part makes you galactically stupid.
1989	78,79,80,81	And that would have been the best course of action. Pit bull owners pull this crap all the time.
1990	13,14,15,16,17,18,19,20,21	There ya go. Stupidity to the max.
1991		Sanders has balls unlike Treyson Goowdy.The Gowdy is a blow hard, has done nothing.What a loser
1992		It's really hard to say this since there are so many, but this might be the dumbest thing you've ever said.
1993	52,53,54,55,56,57,58,59,60	Saying the strategy will work quickly is epitome of stupidity.
1994	30,31,32,33,34,35	Trump Bunnies: You can't fix stupid. Fad, proven over and over every day.
1995	4,5,6,7,8,70,71,72,73,74,75,76,77,172,173,174,175,176	hey loser change your name to something more accurate, as in alt left ignorant for life!,, unlike the dems who take their money from soros or wall street as well,, save it loser you lose again. this is about fixing the mess obamacare created
1996	23,24,25,26,27	And you are a complete moron who obviously doesn't know the meaning of the word narcissist. By the way your bias is showing.
1997	157,158,159,160,161,162,163,164,165,166	Such vitriol from the left. Who would have thought these open minded, tolerant people could be so vile and hateful? It so exposes you all for what you are, hypocrites.
1998		It is now time for most of you to expand your pebble minds.
1999	828,829,830,831	Why does this author think she can demand, or is owed anything from either of these two people? One guy is a goon, the other is illiterate. They aren't law makers, teachers, or in any kind moral authority position. They are entertainers who get punched for her pleasure, and will likely live out their days mentally debilitated from the repeated blows to the head. Do we get to comb deeply through this authors personal history and determine all the groups she owes apologies or explanations to? Why not? As an opinion maker in a national news paper and instructor of young people, she has far, far more influence on Canadians than two ignorant punches. The arrogance of these pseudo-intellectual academics is astounding. Since they are so enlightened and pure, YOU owe THEM an explanation and an apology as to why you're so dumb and ignorant.

Fig. 1. Screenshot of the dataset

B. Data Preprocessing

To transform the dataset into the appropriate format for model training, I had to perform a series of data preprocessing on the test, train and validation dataset. First, I generated for each record in the dataset, a sequence of 1's and 0's

representing the toxicity of each word token in the sentence. For example the sentence "Your comment is ridiculous" will have a sequence of [0, 0, 0, 1] to indicate that the ridiculous word is toxic and the others are non-toxic. I also converted the sentences to lower case and removed the punctuation's marks before transforming the each word and toxic label in the sentence to a row as shown in Fig 3. I then created a out of vocabulary (OOV) embeddings using the Glove embedding vector.

	file_id	sentence_id	token	tag
0	0	0	Another	Non Toxic
1	0	0	violent	Toxic
2	0	0	and	Toxic
3	0	0	aggressive	Toxic
4	0	0	immigrant	Toxic
5	0	0	killing	Non Toxic
6	0	0	a	Non Toxic
7	0	0	innocent	Non Toxic
8	0	0	and	Toxic
9	0	0	intelligent	Non Toxic
10	0	0	US	Non Toxic
11	0	0	Citizen	Non Toxic
12	0	0	Sarcasm	Non Toxic

Fig. 2. Screenshot of the dataset

C. Model Design

The model used to tackle the toxic span detection was Bidirectional GRU RNN. The layer of the network (shown in 3) was consisted of an embedding layer, GRU layer, a fully connected layer and lastly the softmax layer for the prediction.

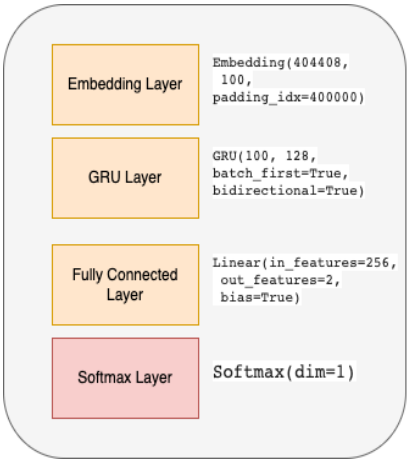

Fig. 3. Summary of the GRU model layers

The model was developed using PyTorch framework and it was trained on google colab platform with GPU runtime. The model was trained for 15 epochs with 2 learning rates (0.001, and 0.003), and 3 different batch sizes (1, 16, and 128).

IV. Experiments

After training the model with the different learning rates, batch sizes and alpha values, the best model was trained with the configuration shown in Table I.

The training validation accuracy and loss plot is shown in Fig 4 and 5 respectively.

TABLE I
Model Training Configuration Summary.

Hyperparameter	Value
alpha	10
batch_size	16
lr	0.001
min_delta	-0.005
validation accuracy	0.937
validation loss	0.238

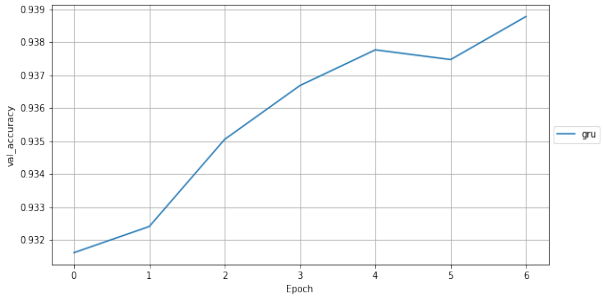

Fig. 4. Validation Accuracy Plot

After the training, the model was used for prediction on the test dataset. The classification report of the model on the test set is shown in Fig 6. The F1 score was 0.98 and 0.59 for the Non toxic and toxic label respectively.

After the prediction from the model, I merged each token into a sentence based on its sentence id and obtained their corresponding actual toxic word and predicted word before getting the index word in the sentence. A screenshot of the final predictions is shown in Fig 7 and Fig 8

V. Conclusion

Toxic span detection is an advanced way of moderating comments and messages posted on online platforms. It goes beyond just classifying a sentence as toxic or not but it is able to extract the exact words that are toxic in the sentence. This is helpful for online platform moderators to ensure that negative and hurtful words are censored in the online space. In this project, I was able to apply a similar approach used in part of speech to toxic span detection by classifying the words in the sentence as toxic or nontoxic. I used the Bidirectional

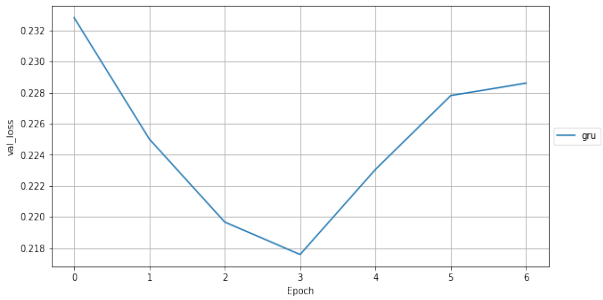

Fig. 5. Validation Loss Plot

	precision	recall	f1-score	support
Non Toxic	0.99	0.98	0.98	62864
Toxic	0.54	0.65	0.59	2453
accuracy			0.97	65317
macro avg	0.76	0.81	0.79	65317
eighted avg	0.97	0.97	0.97	65317

Fig. 6. Classification Report

GRU network and was able to obtain a validation accuracy score of 0.937.

REFERENCES

[1] "Sequence labeling," Wikipedia, 27-Dec-2020. [Online]. Available: https://en.wikipedia.org/wiki/Sequence_labeling. [Accessed: 21-Dec-2022].

[2] EnricoBenedetti, "Pos-tagging: Part of speech tagging using different rnns," GitHub. [Online]. Available: https://github.com/EnricoBenedetti/pos-tagging. [Accessed: 21-Dec-2022].

[3] J. Pavlopoulos, J. Sorensen, L. Laugier, and I. Androutsopoulos, "Semeval-2021 task 5: Toxic spans detection," Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021.

```
df.head(10)
```

	sentence	toxic	predicted	toxic_span	pred_toxic_span
0	thats right they are not normal and i am start...	abnormal,sexist,rubbish	sexist,rubbish	[81, 82, 83, 84, 85, 86, 87, 88, 127, 128, 129...	[127, 128, 129, 130, 131, 132, 134, 135, 136, ...
1	watch people die from taking away their health...	stupid	stupid	[76, 77, 78, 79, 80, 81]	[76, 77, 78, 79, 80, 81]
2	tens years ago i contacted the pdr and suggest...		nothing,but,ignorant,people	[]	[456, 457, 458, 459, 460, 461, 462, 464, 465, ...
3	the parallels between the anc and the sicilian...		ignorant	[]	[400, 401, 402, 403, 404, 405, 406, 407]
4	intel community how can we work for a presiden...			[]	[]
5	i live in an area that saw major flood damage ...	stupid	stupid	[126, 127, 128, 129, 130, 131]	[126, 127, 128, 129, 130, 131]
6	elaine marie jeffers sounds like a sociopath	sociopath	sociopath	[35, 36, 37, 38, 39, 40, 41, 42, 43]	[35, 36, 37, 38, 39, 40, 41, 42, 43]
7	a rock has more mental horsepower than the ind...	ignorant,dishonest,people	ignorant	[55, 56, 57, 58, 59, 60, 61, 62, 111, 112, 113...	[55, 56, 57, 58, 59, 60, 61, 62]
8	ha ha hillary lost all you hillary fools need ...	fools,dump,trash	morons	[35, 36, 37, 38, 39, 152, 153, 154, 155, 198, ...	[306, 307, 308, 309, 310, 311]
9	yesit is to be expectedoddly enough it happens...	dumb	mouth,dumb	[86, 87, 88, 89]	[71, 72, 73, 74, 75, 86, 87, 88, 89]

Fig. 7. Test Result 1

```
df.tail(10)
```

	sentence	toxic	predicted	toxic_span	pred_toxic_span
1990	there ya go stupidity to the max	stupidity	stupidity	[12, 13, 14, 15, 16, 17, 18, 19, 20]	[12, 13, 14, 15, 16, 17, 18, 19, 20]
1991	sanders has balls unlike treyson goowdythe gow...		loser	[]	[87, 88, 89, 90, 91]
1992	its really hard to say this since there are so...		dumbest	[]	[74, 75, 76, 77, 78, 79, 80]
1993	saying the strategy will work quickly is epito...	stupidity	stupidity	[52, 53, 54, 55, 56, 57, 58, 59, 60]	[52, 53, 54, 55, 56, 57, 58, 59, 60]
1994	trump bunnies you cant fix stupid fact proven ...	stupid	stupid	[27, 28, 29, 30, 31, 32]	[27, 28, 29, 30, 31, 32]
1995	hey loser change your name to something more a...	loser,ignorant,loser	loser,ignorant,obamacare	[4, 5, 6, 7, 8, 69, 70, 71, 72, 73, 74, 75, 76...	[4, 5, 6, 7, 8, 69, 70, 71, 72, 73, 74, 75, 76...
1996	and you are a complete moron who obviously doe...	moron	moron	[23, 24, 25, 26, 27]	[23, 24, 25, 26, 27]
1997	such vitriol from the left who would have thou...	hypocrites	hypocrites	[152, 153, 154, 155, 156, 157, 158, 159, 160, ...	[152, 153, 154, 155, 156, 157, 158, 159, 160, ...
1998	it is now time for most of you to expand your ...		minds	[]	[53, 54, 55, 56, 57]
1999	why does this author think she can demand or i...	dumb	dumb,ignorant	[807, 808, 809, 810]	[807, 808, 809, 810, 622, 623, 624, 625, 626, ...

Fig. 8. Test Result 2