COSC420 Borderline Text Classification

Alec Fraser | 4696790

Overview

In this assignment the idea was to classify and predict text for borderline content. Borderline content is close to being banned but not quite enough. Any perspective borderline is very subjective as to where this arbitrary line fits in its respective content. The idea of borderline content is very prevalent in text for the likes oh hate speech and unsavoury messages which are very common on social media if left unfiltered. For my assignment I decided to implement a borderline classifier on a dataset focused on sexist comments. The dataset consisted of online messages and a marking of true or false depending on if the message was determined to be sexist.

Dataset

The dataset I used was obtained from the hate speech corpora called the 'Call me sexist, but' sexism dataset. This dataset can be found at https://doi.org/10.7802/2251, the text that was analysed was directly from twitter and is given a toxicity score and a sexist label. For the purpose of my assignment, I focused on the message and its sexist tag. The toxicity score is so variable on the message it does not give us any useful diagnostics on the data.

The sexism dataset comes in the form of a .csv file, I personally edited the file to remove the ID tags and edit the headers to be in line with my data frame. The column of toxicity was also removed from the dataset as it was unnecessary for my project. The dataset was then split into training and test subsets at a ratio of 80:20.

Classifier

For creating the classifier, I followed the tutorial on text classification with hugging face transformers in TensorFlow on Towards Data Science which was provided in the assignment brief. This tutorial was very insightful on how to use the ktrain library to create and use a classifier and learner. To code and run my classifier I decided to use Google Colab which is a cloud coding service that allows the user to use cloud GPU's for processing. This is a pivotal component in being able to get efficient run times for the classifier to learn from the dataset over multiple epochs.

The transformer model I implemented was the DistilBERT model, which is the distilled version of BERT. DistilBERT was the most ideal transformer to use as for one it can be easily implemented into TensorFlow and is a lighter and faster version of the regular BERT transformer. The fact that DistilBERT is cheaper in terms of hardware usage is the core reason that I used the model as run times are a major limiting factor for my assignment. DistilBERT has been evaluated to have 60% faster while retaining 97% of the accuracy from BERT which is a welcome trade-off as the speed outweighs the accuracy drop.

```
MODEL_NAME =  'distilbert-base-uncased'

categories = ['False', 'True']

t = text.Transformer(MODEL_NAME, maxlen=200, classes= categories)
```

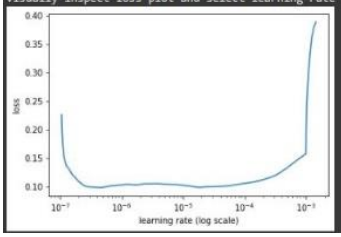
Figure 1 classifier model and transformer

The categories that my transformer classified was the same as the dataset tags of true and false, this keeps the predictor streamlined with the data. The max length I used for the transformer was 200 as this gave a solid training accuracy while preventing a high loss, higher lengths became redundant as the increase in accuracy was minimal and drastically increased training times.

Once my classifier was created, I first trained the model to determine the learning rate of the classifier. This was done by the following process. This found a stable learning rate that minimised loss to be between 10e-4 and 10e-5 so I chose a learning rate of 5e-5 to being in the middle of the two.

```
simulating training for different learning rates... this may take a few moments...
Epoch 1/2
1817/1817 [==============================] - 263s 144ms/step - loss: 0.1644 - accuracy: 0.9395
Epoch 2/2
1817/1817 [==============================] - 13s 7ms/step - loss: 0.4197 - accuracy: 0.8707


done.
Visually inspect loss plot and select learning rate associated with falling loss
```



The model was then fitted over 5 epochs to produce the predictor used for my sexism data. As we can see the accuracy is very high being above 0.90 and over each epoch the overall loss is being lowered which indicated learning is being done and the model is improving over epochs. By the 5th epoch my classifier had a stable validation accuracy of 0.9351.

```
begin training using onecycle policy with max lr of 5e-05...
Epoch 1/5
1818/1818 [==============================] - 290s 149ms/step - loss: 0.2228 - accuracy: 0.9047 - val_loss: 0.1799 - val_accuracy: 0.9091
Epoch 2/5
1818/1818 [==============================] - 270s 148ms/step - loss: 0.1558 - accuracy: 0.9309 - val_loss: 0.1749 - val_accuracy: 0.9124
Epoch 3/5
1818/1818 [==============================] - 276s 151ms/step - loss: 0.1244 - accuracy: 0.9472 - val_loss: 0.1647 - val_accuracy: 0.9267
Epoch 4/5
1818/1818 [==============================] - 271s 148ms/step - loss: 0.0683 - accuracy: 0.9736 - val_loss: 0.1899 - val_accuracy: 0.9351
Epoch 5/5
1818/1818 [==============================] - 273s 149ms/step - loss: 0.0245 - accuracy: 0.9910 - val_loss: 0.2453 - val_accuracy: 0.9351
<keras.callbacks.History at 0x7f217da0fa10>
```

Borderline sexist content

The definition of sexism is prejudice, stereotyping, or discrimination, typically against women, on the basis of sex. When looking at whether a message was sexist or not I tried to base it off if the targeted message was focused on a persons gender as the reason for being hateful. For example, a message that is saying that a person is bad based of if they are female would be considered sexist whereas a message hating someone for a any specific reason but happens the gender is an afterthought would just be considered general hate not based on being sexist. This however is a grey area as with any hateful speech there is no right or wrong and everything can be sexist to a different degree. By this premise I tried to create my borderline to predict in my classifier. For example, a borderline message would be one that is hateful / discriminatory but is vague if the persons gender is the reason as to why the message was considered hateful. This will then cover comments the are negligent or distasteful where there are aspects of the comment being sexist but not enough to blatantly call the comment intently sexist.

The way that I implemented this process into my classifier was through the use of probabilities. When the predictor processes a message a probability for true and false is returned to determine what percent the predictor thinks an individual comment is sexist or not. I used this probability as the core for classing messages that are borderline. I used the probability of a message being classed as true to determine my borderline. A probability of 0 would mean the message is not considered sexist at all going up to 1 which is very sexist. Between these two extremes there is a range which I would classify as borderline. I firstly tried a range of 0.55 to 0.80 for the borderline content with a lower probability being false and above true. This instantly was too high as comments I would still consider sexist were flagged as borderline.

text	toxicity	sexist	predicted_sexist	borderline_predict
Why are men so afraid of housework? Is it because a mop looks like a giant penis?	0.87222266	false	True	Borderline
Soccer is 1 of two sports (tennis) where the women's version is not painful to watch due to inferior athletes. im right	0.4450528	true	True	Borderline
I think all women should be in the 49kg weight division, sports or not.	0.55293214	true	True	Borderline
MENTION1342 No one watches women's basketball I mean seriously not a lot of people do. So the pay is go… https://t.co/m9DD9njlak	0.30566385	true	True	Borderline
If you don't like doing nice things for your female significant other and find it a chore to do once a year, you probably don't love them.	0.19236746	false	True	Borderline

For example ID 231 with the message “I think all women should be in the 49kg weight division. Sports or not..” this clearly is a sexist remark and I believe this too extreme to be classified as borderline. To counteract this, I lowered the max probability down to 0.70 which produced better results.

text	sexist	predicted_sexist	borderline_predict
MENTION2263 They'll if you look at polls it's generally "women voters" that have made up their minds… https://t.co/3AqTyRRcVE	false	True	Borderline
Act like a lady .. think like a man.. https://t.co/KpZhMU7JIv	true	True	Borderline
it's harder for me to be mean to girls than it is for me to be mean to boys	false	True	Borderline
a game with a female as main character sucks	true	True	Borderline
Women are much happier if they stay at home and take care of their children	true	True	Borderline

Discussion

Borderline content is a topic that is inherently very hard to define and give clear parameters as everyone has their own view and weighting to how sexist a comment can be. What someone considers sexist may not be considered with the same gravity for another person. This can cause problems with data collections as different aspects that determine a message may have different weightings across sources, creating data inconsistencies. Yet if there was only one source the data could also be biased. This causes issues when determining if the dataset is a good representation of whether something is sexist or not due to a biased dataset creating a biased classifier.

The idea of something being borderline in of itself is a tough area in terms of management. If something can either be allowed or banned, then if it is on the borderline its in this grey area where its allowed and banned at the same time. This grey area causes major moderation problems as its not going against protocols that have been defined but also can be offensive. If a moderator were to remove the message it would be an infringement of someone’s free speech as they still have adhered to the companies’ policies. Ideally there should be some form of compromise for borderline content, messages that a flagged as borderline should be moderated and be given warnings. If repeated abusive of borderline content is used, then further action should be taken as its going from unintended or distasteful into being purposeful and trying to use borderline content as a loophole.

Ideally this moderation will come as a form of majority vote if moderators determine the message to be sexist with a similar track record, then action can be taken. I personally believe this is an ideal execution of borderline moderation but comes at a heavy cost of manpower to a company and would be hard to manage the amount of data that is sent and received with larger social media platforms like Facebook and Twitter.

References

Samory, M., Sen, I., Kohne, J., Flöck, F. and Wagner, C., 2021, May. Call me sexist, but…: Revisiting sexism detection using psychological scales and adversarial samples. In Intl AAAI Conf. Web and Social Media (pp. 573-584)

Arun Maiya, Jan 15 2021 , Text Classification with Hugging Face Transformers in TensorFlow 2 (Without Tears) https://towardsdatascience.com/text-classification-with-hugging-face-transformers-in-tensorflow-2-without-tears-ee50e4f3e7ed