English Sarcasm Detection using the RoBERTa Report
Akhil Tadiparthi

Problem Space

My problem revolves around an English sarcasm detection model that, when given a text from the dataset, predicts whether or not it is sarcastic. For that, I will be using RoBERTa, a type of deep learning model known as a transformer, which is a type of neural network architecture that has been shown to be particularly effective for natural language processing tasks. RoBERTa was pre-trained on a large corpus of text data and fine-tuned on the iSarcasmEval dataset, which enabled it to learn to identify the linguistic features associated with sarcasm in English. I will be using this model to help us predict a text whether or not it is sarcastic and assign a value to it(1 if sarcastic and 0 if not sarcastic). For example, I can have sarcastic text: “Gotta love people who follow you and unfollow because you don’t follow them within an hour or 2. Sorry, I don’t stay on Twitter 24/7”; for this sentence, the model would assign a predicted label of 1 to indicate that the text is sarcastic. On the other hand, for non-sarcastic text: “I dislike people who follow me, only to unfollow me when I don’t follow back right away. I’m not on Twitter that much to follow right away”; for this sentence, the model would assign a predicted label of 0 to indicate that the text is not sarcastic. This approach to machine learning is known as supervised learning, where a model is trained on labeled examples to learn to predict the correct output for new, unseen examples. In the case of sarcasm detection, the labeled examples are sentences labeled as either sarcastic or not, and the goal of the model is to learn to predict the correct label for new, unseen sentences. This type of model can be an example of how machine learning can be used to automate the task of identifying sarcasm in natural language text. Specifically, for companies such as Twitter, Google, etc. that process a lot of languages through their services, it can be particularly helpful to use models like this to provide information on tweets, posts that are posted by recognized individuals, influencers, etc. My implementation of the RoBERTa model is a little bit different in the sense that I pretrained the data before the actual training phase. Just from this, we significantly improved the F1 score: 0.3 to 0.43

Approach

We had originally used the BERT model initialized with the bert-base checkpoint rather than the RoBERTa model, and only the provided iSarcasmEval Task A training data as our training data, but we could not get an F1-score higher than the 0.30 range. For this reason, we decided to change the model and acquire more training data in order to increase the F1 score. We also experimented using the Glove model and Bigram model but weren’t able to get a good score either. For that reason, we chose to go ahead with the RoBERTa model since it was giving the highest F1 scores. In the data preprocessing step, I remove data irrelevant to sarcasm detection, such as URLs, HTML tags, floats, and Twitter handles, as well as convert all text to lowercase. In addition to the provided iSarcasmEval Task A training data, I used the training and test data sets from SEM 2018 as additional training data. This allowed us to achieve a higher

F1 score. Furthermore, I implemented a classifier class using the RoBERTa model. The classifier has two components: The RoBERTa model and a feed-forward neural network. For the neural network, there are 2 dropout layers with a dropout rate of 0.3, 2 linear layers, and one normalization layer. The activation function used is tanh(x). The input is first passed through the first dropout layer, then the first linear layer, then the normalization layer, passed through the activation function tanh(x), the output of which is then passed through the second dropout layer, and finally passed through the second linear layer. For the pretraining phase, the pretraining function initializes the RoBERTa model using the provided training and validation iterators. Next, The train function trains the model using the same iterator and performs validation at specified intervals. The training parameters I used are detailed in the diagram below:

Training Parameters	Pre-training Loop	Training Loop
Number of Epochs	4	12
Learning Rate	1e-4 = 0.0001	1e-5 = 0.00001
Batch Size	16	16

After testing the model using the test data, the evaluation metric for this task is the F1-score of the sarcastic class, which is computed according to the following formula:

$$F_1^{sarcastic} = 2 \cdot \frac{P^{sarcastic} \cdot R^{sarcastic}}{P^{sarcastic} + R^{sarcastic}},$$

Essentially the same F1 score we have been learning in class, but we are interested in the F1 score for the accuracy of the sarcasm classifications. The results of this against the test data are detailed below in the results section of the report.

Data

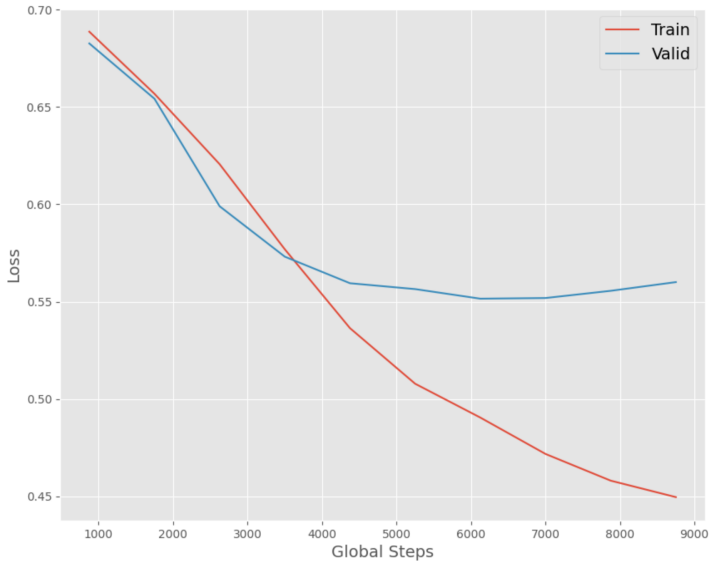
The original training data I got from iSarcasmEval to perform the task was very skewed towards Not Sarcastic texts, with 2601 Not Sarcastic texts and only 867 Sarcastic ones. However, this made the initial iteration of the model to not be as accurate. To address this, we used previous data from SemEval 2018, both training and testing, as well as the training data for SemEval 2022. This dataset had a better balance of the sample set having 10478 sarcastic samples and 9508 non-sarcastic samples. One benefit of this dataset is that the dataset had data augmentation to make the data more evenly distributed, this helped with our f1 score significantly. The features that were used by the model are character n-grams, which involve extracting contiguous sequences of characters of length “n” from the text, which can capture the subtle nuances of language that may not be apparent from individual words. In addition, we have

Part-of-speech tags which use the part-of-speech tags of the words in the text to capture the syntactic structure of the sentence. For example, a sentence with a negative sentiment may be more likely to be sarcastic if it contains a positive adjective. Lastly, we have Emoticons, which can provide a strong signal for sarcasm, as they often convey a tone that is opposite to the literal meaning of the words. Using all this, when we use testing data against the model, we get interesting results outlined below.

Results

	precision	recall	f1-score	support
0	0.91	0.85	0.88	1200
1	0.36	0.53	0.43	200
accuracy			0.80	1400
macro avg	0.64	0.69	0.66	1400
weighted avg	0.84	0.80	0.82	1400

Based on the results we see to the left, can see that our F1 score for determining sarcasm was 0.43. This is a little above average when we compare it to other projects that simply just use the RoBERTa. However, since we first pretrained the data and implemented a neural network, our F1 score is much higher than the average. We also see that our accuracy was 80% which is decent for the model. As for the F1 score of the model predicting the non-sarcasm correctly, we have a 0.88, which is higher than for the sarcasm. We take the weighted average of these two depending on the support to the right of the F1 scores to get the weighted average of 0.82 for both classifications. If we look at the loss curve to the bottom left



we see that when we compare the training and validation loss, we see that as we take more steps, we may be prone to overfitting since the plot of training loss continues to decrease with experience. However, since the plot of validation loss decreases to a point of stability, it would've been a better fit if the training loss has a small gap with the valid loss as it decreases to a point of stability. So, if we target the overfitting, the model would have a much better fit, and proven a good model. I also noticed a couple of places where the model fails. Specifically, when it comes to really long sentences the model automatically predicts that it's not sarcasm. As seen by the diagram below on the next page, the model is

classifying all those long sentences as a 0(not sarcastic) even though the actual is a 1(sarcastic).

241	I'm heading out to snag my COVID-19 inoculation! Thanks government mandate! Super pleased to be doing my part even though the vaccines do not stop one from getting the disease, they do not stop the spread, and it leads to a life of non-stop useless boosters! Woo-hoo!	1	0
488	I went to the best restaurant today for dinner. It was absolutely amazing. My steak was cold and the fries were rock solid. I expected my food to be mouth watering and it sure was when I had to wash each bite down with my drink.	1	0
644	Why does Ramish think he can stop my kitchen rennovation because of the noise? His kids are not exactly quiet or competent when they play their musical instruments, lets just say my dog makes a better sound.	1	0
724	It's actually really easy to get into the mindset of a Tory. All I do is smash my head into a wall until I bleed and then slur incoherently about social mobility and England before stuffing handfuls of loose change into Tim Martin's hands and kick a homeless person in the jaw.	1	0
779	You were given the option to be born into a more financially stable family at conception, but chose not to avail yourself of the opportunity provided; perhaps you should now take personal responsibility for that decision, and deal with the consequences?' 🎩	1	0
812	My sisters kid, when we were playing in the living room, told me that I look like ghoul due to my baggy eyes. I love children so much, they make me feel so appreciated and loved, always say the kindest things!	1	0
872	I'm so glad that smart people are still refusing to wear masks in shops. Thanks to their selflessness it highlights that if we just ignore something it'll go away – don't know why we haven't all tried it.	1	0
931	Hi John, thank you SO MUCH :-) for the text last night. ordering (!) me back to work off holiday for this morning as someone else is sick. Sooooo Sorry :-(that I am in Beijing with an 8 hours time difference and couldn't make it! I'll try harder next time.....	1	0
995	School sends out a newsletter* So now I need a whiteboard to track ALL the school festive activities!! Mufti day - don't forget to make a payment for this, carol concert, early school closure etc Does it ever stop!	1	0
1011	I think women should be able to join men's sports teams and vice versa. I really don't see any problem with for example having a 5 foot 3 inches slim women versus a 300 pound heavyweight boxer because i feel like women are very much equal to any man. Even much faster and stronger men. If anybody takes offense to my point then they need to reevaluate their lives as their are no man that could do anything better than a lady.	1	0
1084	I'd like to thank my current employer for making me feel accepted and valued. It's particularly nice that, whenever I have some suggestions for them, that they seriously consider them and don't just automatically put their fingers in their ears and shout 'I CAN'T HEAR YOU'. Metaphorically.	1	0
1150	My favourite thing driving to work in the morning is how peaceful and slow everyone is. I love when people don't know where they're going, like they are just out for a stroll at 8:30 in the morning. It makes me so happy.	1	0
1186	Top 10 pools in my book: 1. Swimming pool 2. Paddling pool 3. Above-ground pool 4. Family pool 5. Architectural pool 6. Indoor pool 7. Lap pool 8. Olympic size pool 9. Natural pool 10. Salt water pool Sorry Liverpool you are not top 10 pools in my book 😎😎😎	1	0
1204	The omicron variant has decided to give us all a wonderful gift for Christmas. For that one day it has chosen to not be spread between people at all. This gift will make me feel a lot better about being around loved ones who still refuse to get the vaccine or to wear a mask at the very least.	1	0

This shows that the model might need to be reworked around that to correctly classify long sentences. Another area where the model is failing significantly is with the use of emojis in the sample. As seen by the picture to the left, we see that the model always classifies the samples with emojis as a 1(sarcastic) even though the samples should be classified as a 0(not sarcastic). Other than that, most of the classifications are generally other than tweaking the model to make it more accurate for the future.

Just as we all get told to work from home again! 🤫	0	1
How are you not even going to have a look and review that!? Absolute joke	0	1
o the omincron variant and can't wait to go shopping whilst wearing a facemask.	1	0
Officially promoted to General Manager 😁 Today has been a good news day	0	1
Today, I lost a follower because I said I was unwell.\n\nDontcha love Twitter.	0	1
t today would be friday.....wow, guess what happened...\nDreams are amazing!	0	1
Omicron: Cases are going up, run for the hills.	1	0
Woke up and found £10, today will be a good day	0	1
ıo crimbo! The government designed this covid variant just in time for christmas!	1	0
You on your best behaviour is me on my worst	1	0
how have i just got an unconditional offer at york, so happy🥰	0	1
ıake a payment for this, carol concert, early school closure etc Does it ever stop!	1	0
manual for schools to hand out?. Some parents could really use the guidance…	1	0
ıd to stock up on office supplies at Tesco but all I could find in the aisle was wine	1	0
who phone up Jeremy Vine on 5. Everyone one of them has a budgie in a cage	0	1
k lockdown now to keep up the pretence that this is about public health not cash	0	1
BBC reporting on daily covid cases now how convenient🤣	0	1

Discussion

A future direction to take this model would be to experiment this model with other languages so that we expand these capabilities across different languages. Current research has been studying Arabic but soon enough more languages will be added to the list to potentially create accurate models for sarcasm detection. For my project, I only scratched the surface, which is to simply classify whether a text is sarcastic or not. However, we can take this model in many different, more complex ways in the future. For example, we can add a binary multi-label classification task where, given a text, the model will determine which ironic speech category that sample belongs to, if any. Another area that the model can expand to is when a sarcastic text and its non-sarcastic rephrase(two texts that convey the same meaning)

are given, the model should be able to determine the actual sarcastic phrase. Lastly, we can also simply combine other newer developed models with RoBERTa to possibly make a more accurate model with higher accuracy scores.