AI Cracks Sarcasm Code

Oğuz Baz, Nina Koperska, Aleksandra Naumova

University of Amsterdam

Introduction

Our project aims to shed light on the comparative strengths of LSTM and BERT in sarcasm detection within news headlines. Transformers have excelled in text analysis tasks using bidirectional context, efficient parallelization, attention mechanisms, and pretrained representations. However, there has been some findings suggested that bidirectional LSTM's can achieve better results than BERT on small datasets and the simple models are usually trained in less time. Therefore, the performance of a model can be dependent on the task and the data. For that reason, we are going to split the data into short and long headlines and analyse the performances of LSTM and BERT models in this project.

Dataset

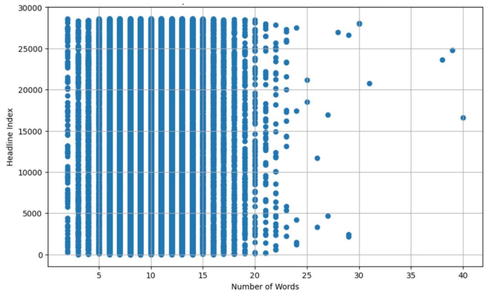
DATA DESCRIPTION

News Headlines Dataset contains headlines from two news websites: "The Onion" aims at producing sarcastic versions of current events (14K headlines), whereas "HuffPost" publishes real news (15K headlines).

Link: https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

DATA SPLIT

To test the performance of models on short and long sentences, the original dataset had to be split based on sentence length. Upon visual inspection of the distribution of data (see figure below), potential outliers were identified, and finally excluded if their lengths exceeded 25 words. Afterwards, data was split according to the median, which resulted in a shorter sentences dataset with less or equal to10 words, and longer sentences with more than 10 words and less or equal to 25 words. An equal distribution of sarcastic and non-sarcastic headlines was also considered during the split into these two conditions.



Methods

To check whether there is a difference in performance of LSTM and BERT on long and short sentences, four models were run on the long and short sentences datasets:

Experimental conditions

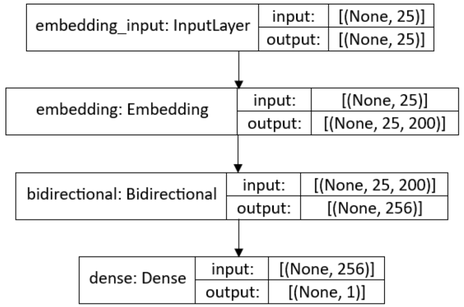
BERT + short sentences	BERT + long sentences	LSTM + short sentences	LSTM + long sentences

Bidirectional LSTM

Bidirectional Long Short-Term Memory (LSTM) is a type of recurrent neural network architecture used in natural language processing. Unlike traditional LSTMs, bidirectional LSTMs process input sequences in both forward and backward directions, enabling the model to capture information from both past and future context, which makes it well-suited for the tasks of sarcasm detection.

ARCHITECTURE

Two models were built to handle long and short headlines separately. The only difference between the models is the input length in the Embedding layer, which results from different padding - short headlines were padded to a length of 10 words, while long headlines were padded to a length of 25 words.

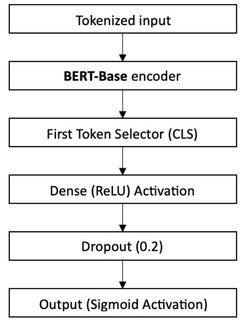


Compiling & Training

	Optimizer	LR	Loss function	Batch Size	Epochs
LSTM	Adam	0.01	Binary cross-entropy	128	2
BERT	Adam	1e-5	Binary cross-entropy	32	5

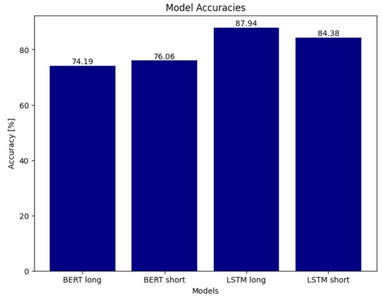
BERT

Bidirectional Encoder Representations from Transformers (BERT) is a machine learning model for language processing which can be used for tasks like sentiment analysis, text prediction, and text generation. The advantage of BERT is that (1) it considers both directions of the text input, (2) it is pretrained on masked language modelling and next sentence prediction, which helps it learn complex language patterns.



Results

- LSTM performed better on both short and long datasets, with accuracies of 84% and 89% respectively
- LSTM performed much better on long sentences data, with stable validation accuracy on both datasets, while training accuracy continued increasing
- BERT performed slightly better on the long sentences data, and similarly to LSTM its validation accuracy was initially much higher than the training accuracy, but then training accuracy picked up while the increase in validation accuracy remained stagnant



Discussion & Conclusion

The aim of this study was to answer two questions:

1. Is there a difference in performance on the chosen dataset between the LSTM and BERT models?
2. 2. Is there a difference in performance of the two models on shorter versus longer input sentences?

The results found however, only partially managed to answer these questions. LSTM performed better than BERT on both datasets, but the results were too inconclusive to make any inference regarding possible differences between short and long sentences. It is important to note that the lower accuracy from the BERT model found in this project could be due to insufficient hyperparameter tuning. It could also be, that BERT Base doesn't perform as well as LSTM on shorter inputs like headlines in this case. This matter requires further investigation and contrasting the performance of BERT with LSTM on other datasets. Another point is one relating to the experimental design. Splitting data into short and long sentences was an arbitrary choice, and it could be the case that there are differences in how LSTMs and Transformers perform on texts of varying lengths, but they were just not visible on our data due to little qualitative differences between conditions. A suggestion for future research would be to test the performance of the two models on longer texts, to confirm whether there is indeed a lack of support of this hypothesis.

References

- A. Zeyer, P. Bahar, K. Irie, R. Schlüter and H. Ney, "A Comparison of Transformer and LSTM Encoder Decoder Models for ASR," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 2019, pp. 8-15, doi: 10.1109/ASRU46091.2019.9004025.
- Ezen-Can, A. (2020). A comparison of LSTM and BERT for small corpus. arXiv (Cornell University). https://www.arxiv.org/pdf/2009.05451Sun, C., Qiu, X., Xu, Y., &
- Huang, X. (2019). How to Fine-Tune BERT for text classification? In Lecture Notes in Computer Science (pp. 194–206). https://doi.org/10.1007/978-3-030-32381-3_16