

Course Project Progress Report

Chua Yeow Long, ylchua2@illinois.edu
Text Classification Competition

Introduction

For the text classification competition on sarcasm detection, I'll utilise transfer learning using embeddings such as GLoVe and BERT which are both major milestones in modern NLP rather than training a machine learning model from scratch. GLoVe embeddings is one of the first applications of transfer learning in NLP and with the introduction of BERT by Google in 2018, it addresses the shortcomings of LSTMs and other modern NLP techniques and also more recently with huge models such as XLM and GPT-2 achieved state-of-the-art performance.

Methodology & Results

I'll first preprocess the text data by removing stop-words and non-alphabetic characters such as symbols before feeding the data into 3 different models. The first is to use an GLoVe embedding and adding some LSTM and dropout layers and start training the model. A 85% train and 15% validation split is performed to obtain the validation data. For the input text data, I have tried with and without context and only for the BERT case the F1-score is pretty significant. I have tried different LSTM layers and nodes as well as dropout layers and the best f1-score I have obtained is around 0.661. The second method is to use BERT and re-train the entire BERT layers using the dataset on a pretty beefy machine. The best F1-score I have obtained is 0.7378 with context information and 0.716 without context. The baseline F1-score is 0.723 and the top F1-score is by awe with 0.7653.

Description	F1-score
GLoVe Embedding with LSTM layers without context	0.661
GLoVe Embedding with LSTM layers	0.661
BERT without context	0.716
BERT	0.7378

Future-works

I'll try and see if I can get a better score by exploring two different approaches: feature engineering and/or GPT-2/XLM transfer learning based methods when there is time to spare. I have included the references for BERT as well as GPT-2 in the references section below.

References

Language Models are Unsupervised Multitask Learners

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever

Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin