Research Paper Classification by Abstract

Authors: Huandong Chang, Elijah Tamarchenko

# Introduction

In this project, we build a Natural Language Processing (NLP) pipeline and apply LSTM and DistilBERT models to predict which subject a research paper belongs to based on its abstract. Our dataset consists of 15,000 papers from the arxiv dataset on Kaggle, 60% (9,000) of which are used for model training. Paper subjects are uniformly distributed among Computer Science, Mathematics, Physics, Quant. Bio, and Statistics. Our pipeline includes data import, data preparation, model buildings and training, result analysis, and conclusion. With the limited amount of computing power and training data, the LSTM model reaches 86% accuracy and the DistilBERT model reaches 92% accuracy in the testing test.

# Previous Solutions

There are two main approaches for the classification of articles that are used today. There is the metadata approach, which classifies articles based on the metadata which is freely available on the internet, such as title, authors, keywords, general terms, etc. Then there is the content approach, which classifies articles based on the actual content of the research, and is mainly based on NLP classification. Our research falls into the first category, since the abstract is considered part of the metadata, and is something that is freely available for every article online. Classification of articles using metadata is very important, since before buying access to articles researchers want to be sure that the article they are paying for will be useful to them, and thus will not have access to the actual research to run a classifier on it. Results by Yohan et. al. (2014) show that classification of newspaper articles and wiki articles can fall into a range of accuracies of .80-.94%, which provides us a goal range to hit and hopefully improve on. Other work by Kandimalla et. al. (2021) focuses specifically on training models using solely the abstract of the papers, and this was done using an RNN. Their training was done using 9 million data points across 100 different categories. This paper achieves an accuracy on individual categories ranging from 0.50 to 0.94%. Finally, Truong et. al. (2016) used data from 10 different categories of Vietnamese research papers, and was able to achieve an average range of precision of .72 - .91% using different models. However, in our background reading, we did not find mention of using large-scale models like BERT or LSTM models for this specific task, which is what we decided to tackle.

# Dataset

Our dataset is from the arxiv dataset on Kaggle, with over 1.7 million scholarly papers across STEM. Since our computing power is limited, we only extract 15,000 papers from the arxiv dataset and use 60% (9,000) for model training, 20% for the validation set, and 20% for the testing set.

Dataset Link: https://www.kaggle.com/Cornell-University/arxiv?select=arxiv-metadata-oai-snapshot.json

# Proposed Method

## Model 1: Long short-term memory (LSTM)

LSTM is a type of recurrent neural network that can keep track of long-term dependencies in the input sequences. After data preparation and tokenization, we use a sequential model with four layers: a) Embedding; 2) Spatial Dropout 1D; 3) LSTM with 128 units (output dimensionality); 4) Dense layer with softmax activation for subject prediction.

## Model 2: DistilBERT

DistilBERT (Sanh et. al, 2020) is a transformer model based on the BERT architecture. Since the size of the model was a big consideration for us, DistilBERT was the better choice as it is 40% the size of the BERT model and 60% faster, while compromising only 3% on accuracy. This is done using knowledge distillation, compression, and improvements in the linear algebra operations used for computation. Besides this, the model's architecture is the same, with an adjustment to the number of layers and neurons in the model.

# Results and Discussion

Overall, the LSTM Model has given us 86% accuracy and the DistilBERT model gives 92% accuracy in the testing subset. This makes sense, as the DistilBERT model is a finetuning of a large state-of-the-art model which is used for lots of other NLP tasks, and thus would perform a lot better. Also, the use of transformers allows the model to use the context of the abstract for each word in it, rather than sequentially building up from left to right as the LSTM, and this bidirectionality could allow for more accuracy. For both models, Physics paper is the most distinguishable of all the fields (more sensitive and more precise), Math and Quant. Bio have decent performance, while Computer Science and Statistics are harder to identify and are sometimes confused with each other. After analyzing the misclassifications, we find out that a lot of the papers that are misclassified by the model seem to fall under the intersection of computer science and statistics, even though they are only classified as statistics. However, accuracies of 86% and 92% are very impressive for a model that is only reading the abstract. We postulate that further research would show that reading more parts (such as the conclusion part) of each scientific paper would probably provide higher accuracy for either model, as it would provide a lot more context to the transformer model as well as the LSTM model.

# Bibliography

Kandimalla, B., Rohatgi, S., Wu, J., & Giles, C. L. (2021). Large scale subject category classification of scholarly papers with deep attentive neural networks. *Frontiers in research metrics and analytics*, *5*, 31.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Truong, Q. D., Huynh, H. X., & Nguyen, C. N. (2016, March). An abstract-based approach for text classification. In *International Conference on Nature of Computation and Communication* (pp. 237-245). Springer, Cham.

Yohan, P. M., Sasidhar, B., Basha, S. A. H., & Govardhan, A. (2014). Automatic named entity identification and classification using heuristic based approach for telugu. *International Journal of Computer Science Issues (IJCSI)*, *11*(1), 173.