TODO: TITLE: a descriptive title for your term project

Chloe Sheen csheen1@seas
Tina Huang email@domain
Joseph Liu email@domain
Worthan Kwan email@domain
Mia Mansour email@domain

Abstract

TODO: Your abstract should give an overview of your project and your results (100 words).

1 Introduction

TODO: Need to add a figure.

Authorship classification is an essential topic in Natural Language Processing, and it can be used in tasks such as identifying most likely authors of documents, plagiarism checking, and as a new way for recommending authors to readers based on the reader's preferred style of writing. Authors have unique writing styles to their works and are often consistent across a range of different topics and document styles. There have been numerous approaches to handle this task of characterization and classification of authorship, including using a bag-of-words model or Word2Vec. In particular, we aim to experiment with the state-of-the-art BERT pre-training for this classification task.

2 Literature Review

As mentioned, there has been a long history of handling the task of authorship classification. Here, we briefly describe the approaches that are used and how well the approaches worked.

2.1 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., 2018)

This paper introduced the new language representation model Bidirectional Encoder Representations from Transformers (BERT). BERT was designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The pre-trained BERT model can then be fine-tuned with just another output layer to create models for a wide range of tasks, including but not limited to answering questions, language inference, and classification.

The breakthrough of BERT was expanding the fine-tuning approaches of the pre-trained representations to be bidirectional. The authors proposed BERT to alleviate the unidirectionality constraint by using a masked language model pre-training objective that randomly masks some of the tokens from the input with the objective to predict the original vocabulary id of the basked word based on its context. Additionally, they used a next sentence prediction task that jointly pre-trains text-pair representations.

With BERT, the authors were able to demonstrate the merits of bidirectional pre-training of language representations, which was in contrast to the then state-of-the-art unidirectional language models and shallow concatenation of independently trained left-to-right and right-to-left LMs. They also showed that pre-trained representations reduce the need for several heavily-engineered task specific architectures by demonstrating that BERT as the first fine-tuning based model that achieved state-of-the-art performance on a large suite of sentence-level and token-level tasks. Finally, BERT advanced the state-of-the-art for eleven NLP tasks.

2.2 Tweet Classification with BERT in the Field of Disaster Management (Ma, 2019)

This paper applies deep learning techniques to address Tweets classification problems in the disaster management field. The goal is to use disaster-related information for emergency response and better disaster management in the fields of outbreak detection, evacuation study, hazard and damage assessments. Accurate message classification is a necessary requirement to make decisions in the field of disaster management. Effective filtering techniques are needed to filter out noisy information in the user-generated data.

The baseline is the bidirectional LSTM with GloVe Twitter embeddings. The authors tested the default BERT, BERT + nonlinear layers, BERT + convolution, and the BERT-based LSTM which attained the best results. The BERT models are built based on the pytorch-pretrained-BERT repository by huggingface. Five metrics were used to evaluate these models: accuracy, Matthews correlation coefficient, precision, recall, and F-1 score.

That are multiple factors that have lead to misclassifications: (1) ambiguity and subjectivity, (2) lack of context, (3) non-ascii words emojis, (4) semantic misconstrue (sarcasm, metaphors..), (5) keyword influence or misleading hashtag, (6) events that might not happen, (7) short messages, (8) debatable source. If the quality of the data were to be improved, that would boost up the performance of the classifier.

2.3 Deep Learning based Authorship Identification (Qian et al., 2017)

This paper applies deep learning to author classification of a Reuters 50-50 (news) and Gutenberg (story) dataset. The three major models were a GRU (Gated Recurrent Unit) network , LSTM (Long Short Term Memory) network, and a Siamese network. The most impressive models are almost universally deep learning based neural networks. Variation between models comes from the vector representations of the words.

The first dataset is the Reuters 50-50 (called C50) subset of RCV1. RCV1 Is an archive of over 800,000 manually categorized newswire stories. The top 50 authors by total article size were used. Each selected text had at least one subtopic of the class CCAT (corporate/industrial) in order to minimize the odds of classifying by topic accidentally.

The second dataset is the Gutenberg dataset consisting of over 53,000 books. 50 of the most popular 100 authors were chosen. The books cover a wide range of topics and styles in order to minimize the odds of classifying by topic accidentally. Books were edited to remove noise (page numbers, table of contents, information on contributors).

For modeling, GloVe[10] word vectors of size 50 were used on pre trained word embeddings with a total vocabulary of 400,000 tokens. Numbers and special characters were eliminated to allow matching of word representations during parsing.

After optimization, the Article-level GRU performed best on authorship identification with an accuracy of 69.1% on C50 and 89.2% on Gutenberg. The highest testing accuracy of LSTM was 62.7%. The siamese model was used for verification and achieved an accuracy of 99.8% on both datasets.

We will be using this paper to implement as our published baseline. The paper explores the two datasets we are most interested in: Reuters 50-50 and the Gutenberg dataset, along with different deep learning models. Since the publication of this paper was in 2017, and given the introduction of the BERT language representation model in October 2018, we believe that the neural networks implemented in this paper would serve as a good baseline to test the performance of a newer model on the same dataset.

3 Experimental Design

4 Experimental Results

5 Conclusions

TODO: You should write a brief summary of what you accomplished in your term project. Did any of your implementations reach state-of-the-art performance on the task? If not, how close did you come? If not very close, then why not? (100-300 words).

Acknowledgments

TODO If you used someone else's code or you benefited from discussions with one of the TAs, then you should thank them here. Give credit generously! (Optional)

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Guoqin Ma. 2019. Tweets classification with bert in the field of disaster management.

Chen Qian, Ting He, and Rao Zhang. 2017. Deep learning based authorship identification.

A Appendices

This can include short snippets of code that were relevant to your project, along with a description of

what it's doing. It could also include more examples of your training data or your system's output. (Optional)

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here.

LaTeX-specific details: Use `\appendix` before any appendix section to switch the section numbering over to letters.