Project Title

Tao Sun sun955 **Yohan Berg** yberg

Abstract

Authorship verification is the process by which we can determine whether or not someone is the true author of a text. This is not to be confused with authorship attribution, which takes texts and tries to determine which author wrote the texts. Authorship verification is more specifically trained on the writings of a singular author in the attempt to determine whether or not a new writing is written by that specific author. There has been some research done in this field by creating an intrinsic model and taking the problem as a one-class classification problem. This would lead to us training the model based only on text written from our desired author. What we will be looking at and measuring in this paper is creating an extrinsic model that considers papers written by other authors as a negative class and turning this problem into a binary classification problem.

1 Introduction

The topic that we have chosen for our project is Authorship Verification. Within this field, there is relatively little research that measures the effectiveness of different text representation models in verifying authors. Thus, our goal is to offer a comparison between the model's performances based on different forms of text representations. To carry out this experiment, we will take a look at a few text representations and see how they affect the performance of the model. Some of these that we plan on working with are bag of words, term frequency - inverse document frequency (TF-IDF), word embeddings, and contextualized word embeddings. Since many papers and experiments currently evaluate on one or two of these representations, it will be helpful to determine if more complex representations will offer non-trivial improvement to the performance of a model.

To sum up our project, we will attempt to answer whether or not advanced textual representation techniques will significantly improve a model's performance in single-domain or cross-domain authorship verification tasks. This will allow future researchers to determine whether or not the impact of a specific representation is great enough to choose one over the others.

2 Dataset

For our dataset, we will use the PAN 2021 Authorship Verification dataset presented during the PAN 2021 NLP competition (Kestemont et al., 2021). The dataset contains 276,000 textual pairs of transformative literature, or "fanfictions" over a variety of domains such as "Harry Potter," "Marvel," or "Twilight." The cross-domain facet of this dataset would also be helpful in distinguishing our model's effectiveness across similar or distant domains. The dataset also provides three pre-implemented baseline models including a compression-based approach (Halvani and Graner, 2018), naive distance-based model, and a first-order Bag of Words model (Kestemont et al., 2016).

3 Proposed Approach

To answer our research question, we will first conduct data preprocessing to handle any missing or erroneous data as well as analyzing the data for imbalanced labels or domains. Then, we will produce the discussed distinct embeddings of our text input. We will then fit a LSTM and MLP model to the data. Finally, we will evaluate the above models and compare their performance under different text representations to answer our research question.

For the specific embeddings we plan to implement, we will fit a Bag of Words model as our baseline and a TF-IDF model and a word2vec model

for our experimental models. We will also train our experimental models using the pre-trained GloVe embedding (Pennington et al., 2014) and the OpenAI text embedding API.

3.1 Computational Resources

While numerous, our proposed models can be restricted in terms of complexity. Thus, we are not overly concerned about computation resources. We have access to a GPU-powered laptop with 8GB of virtual memory and a similar personal desktop. If computational resources remains an issue we can sample smaller subsets of our data and Purdue IT also offers several options for additional computational resources such as the mc18 server.

4 Evaluation

For each pair of texts, the model should output a probability $p_i \in [0, 1]$ representing the confidence that the two input texts were written by the same author.

In order to evaluate the outcome of our project, we plan to utilize a few error measuring metrics, including F1-score and confusion matrices to measure our accuracy, as well as cross validation to make our measurements more accurate. Since our project's goal is to verify an author's work, we can easily determine false negatives and false positives from true values.

For our baseline, we will use the Bag-of-Words model since Bag-Of-Words is widely regarded as the basic text representation scheme. Additionally, the PAN 2021 dataset we propose to use includes three pre-implemented baseline models which can be used as an overall performance baseline for all models.

We also intend to do some qualitative evaluation. We expect to achieve a decent accuracy F1 score and more insights into how we can accomplish such a task. We will measure if we achieved the intended goals based on the evaluation below.

4.1 F1-Score

The F1-score will allow us to better measure and tune our model based on precision and recall values. Having a balance between these values should lead us to optimal results, allowing us to correctly assess our overall effectiveness in correctly identifying works written by different authors.

4.2 ROC AUC score and Brier Score

To help compare our models to the baselines provided by the PAN 2021 dataset, we will also score our models using the area under the ROC curve and Brier score (BRIER, 1950) implemented using the scikit-learn library.

4.3 Confusion Matrix

A confusion matrix will help us map out which types of errors we have and allows us to prioritize minimizing false positives, false negatives, or both.

4.4 Qualitative Evaluation

Along with both the other evaluation techniques, we will also analyze the types of mistakes the model is making in a qualitative way by comparing precision, recall, and visually looking at output and previous evaluation results to determine patterns between errors.

References

GLENN W. BRIER. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Oren Halvani and Lukas Graner. 2018. Cross-domain authorship attribution based on compression: Notebook for pan at clef 2018. In *Conference and Labs of the Evaluation Forum*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. Overview of the Authorship Verification Task at PAN 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Mike Kestemont, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of julius caesar. *Expert Systems with Applications*, 63:86–96.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 62, New York, NY, USA. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Efstathios Stamatatos. 2016. Authorship verification: A review of recent advances. *Res. Comput. Sci.*, 123(1):9–25.