BERT Fantasy Classification

Justin Widjaja Reiss
Justin.Reiss@colorado.edu

Aman Satya
Aman.Satya@colorado.edu

Introduction

1.1 Problem Statement

This project explores whether using a BERT model on data drawn from a Fantasy domain causes the model to perform differently. Although the BERT model and its iterations are proven to be very effective on complex NLP tasks in GLUE benchmarking, they have not been thoroughly evaluated to see how they work on data from fantasy domains. Since a BERT model is trained using data drawn from the real-world, it could theoretically produce worse results when its embeddings are used to fine-tune on a task that uses fantasy data.

1.2 Objective

This project is focused on analyzing how BERT embeddings interact with fantasy data. By examining these interactions alongside data drawn from a realistic context, we are able to determine whether BERT embeddings can distinguish between fantasy data and realistic data. If BERT embeddings are not significantly influenced by the different context that fantasy data is drawn from, then they could be used on tasks involving fantasy data without significant changes from how the task would be approached in a realistic domain. We analyze these interactions through experiments with genre classifiers, Next Sentence Prediction outputs, and authorship classifiers.

1.3 Scope

Since we pulled our data directly from books, all of the data that we worked with was unlabeled. In response to this limitation, our experiments revolve around examining the BERT embeddings through classification tasks. Although this limits the types of tasks that we can examine using the BERT model, it ultimately does not interfere with the project's goal of examining how BERT interacts with fantasy data. If the embeddings generated by the fantasy data are different, then identifying that difference is important because it could be propagated through the model and produce poor performance on downstream tasks.

1.4 Literature Survey

There isn't any specific research done in the past that examines Fantasy data in this way. However, we referred to various papers which gave us an idea of how to approach the problem statement. Apart from referring to research papers, we understood the basic application and process of how to use BERT models from the Visual Guide to Using BERT from Github (Cited in References Section). The papers that we referred are:

[1] *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

> This paper helped us in understanding the architecture of the BERT model. It is a pretrained model which can be fine-tuned to a task by adding one or more extra layers at the end. This architecture can be used for various kinds of applications such as classification, question answering, and next sentence prediction. In our project we are examining two of these applications: classification and next sentence prediction.

[2] *Multi-Classifier System for Authorship Verification task using Word Embeddings*

> One of the experiments from our project is an authorship classifier where we are classifying the text from various books with respect to its authors. The output from the hidden state embeddings from NSP were used as an input to the LSTM model that we created for classification. This paper gave us an idea of how word embeddings are used for authorship verification tasks.

Experiments

2.1 Datasets and Preprocessing

We used datasets of sentences taken from books on Project Gutenberg. These datasets were chosen to explore various aspects of how BERT interacts with realistic and fantasy datasets. The bird history and current history datasets are used to represent realistic information that would be contained in a book, which helps us avoid comparisons between domains that are too different.

The fantasy books that we chose were *The Wonderful Wizard of Oz* by Lyman Frank Baum, *The Book of Wonder* by Lord Dunsany, *The Legends of King Arthur and His Knights* by Sir James Knowles, *Irish Fairy Tales* by James Stephens, and *The World of Ice and Fire* by George R.R. Martin. The first two fantasy books were chosen to represent books that contain fantastical elements but are otherwise written like other fiction. In contrast to these datasets, the Arthurian Legend and Fairy Tale datasets were chosen because they were written in an older style of English that would distinguish them from modern books. Finally, "The World of Ice and Fire" was

chosen because it is written as a history book for a fantasy world. Using a wide variety of datasets allows us to examine the results with consideration to whether the style of the book is having more of an effect than the fantasy elements themselves. To further examine this effect on a realistic domain, the realistic books were selected to have different types of information and styles of writing. While the bird dataset presents more technical terminology, the current history dataset is written more anecdotally.

The data was processed using the code in Processing.ipynb to split the books into sentences and format the data. Data that is not valuable to the task, such as punctuation, capitalization, and chapter indicators, is removed from the sentences. Once the sentences are properly formatted, they are stored in .csv files corresponding to each dataset alongside a label. These labels are binary values that indicate whether or not a sentence was drawn from a fantasy domain.

2.2 Genre Classification Experiment

	Dorothy	Arthur	BookofWonder	IrishFairy	IceandFire
Bird History	98.33	97.77	94.6	96.66	96.38
Current NYT	93.66	96.83	90.89	92	94.17

Table 1

Implementation

Our first experiment revolved around implementing a genre classifier that determines whether an input sentence is from a realistic or fantasy domain. This classifier uses embeddings of input sentences generated by the BERT model and trains a binary logistic regression classifier to distinguish between realistic and fantasy embeddings. In order to test this classifier, we ran it on combinations of sentences from each realistic dataset with sentences from each fantasy dataset. We ran these tests using the free machines provided by Google Colab. Due to the limited RAM provided by a Colab instance, we compensated for the high RAM usage of BERT embeddings by limiting the number of sentences that we consider to the first 1200 sentences of each dataset, provided the dataset is of that length or longer. The classification accuracies on a 25% test split for each combination of datasets are shown in Table 1.

Analysis

Notably, all of the combinations achieve above a 90% classification accuracy. These results are interesting because they imply that the BERT embeddings contain distinguishable features between the two genres. Additionally, the fact that the combinations with the Arthur dataset, which uses an apparently different style of Old English prose, achieves some of the highest accuracies shows that the style of a

dataset's writing is captured in the model's features. The presence of these features may mean that certain styles of fantasy writing may produce better or worse results than others.

2.3 Next Sentence Prediction Performance

	Bird History	Current NYT	Dorothy	Arthur	BookofWonder	IrishFairy	IceandFire
Loss Values	3.6639	3.672	3.6349	3.77	3.6716	3.614	3.72

Table 2

Implementation

One of BERT's core tasks is a Next Sentence Prediction (NSP) task to learn whether the second sentence in an input actually follows from the first sentence or is a random sentence. We tested whether the BERT model performs well on this task using a variation of the task. This variation uses subsequent sentences from the dataset in the first half of examples as normal, and it uses random sentences drawn from the opposite genre for the second half of the examples. Unlike the genre classifier inputs, these inputs were generated by choosing a number of random first sentences from the datasets that were being tested. After choosing the random first sentences, the corresponding subsequent sentences or random sentences from random datasets were concatenated to these sentences according to the type of example being generated. After formatting the inputs, we passed them into a BERT model wrapper from the python Transformers library that includes an NSP classification head. The losses that the model produced on the NSP task are contained in Table 2.

Analysis

Since the loss functions do not vary by a significant amount per each dataset, it appears that the BERT model can not fit a next sentence prediction task that works across multiple domains of data. This poses a problem when working with fantasy data because it often mixes fantasy elements in with a modern style to make the book more readable.

2.4 Authorship Classifier Experiment

Implementation

An authorship classifier outputs a predicted author from a predefined set of authors for each input. As an extension of the NSP task from above, we used the hidden state embeddings from the BERT NSP task as the inputs into our classifier. This classifier used a series of 32 LSTM units into a Softmax layer that would output predictions across 8 labels. Seven of these labels corresponded to each of the

different datasets that we worked with, and the last label indicates that the two sentences are a combination of one realistic and one fantasy sentence. We added the mixed label to see whether the classifier can identify a mixture of realistic and fantasy elements in a sequence.
Due to the large size of BERT embeddings for two concatenated sentences, we used an instance on Google Cloud to get more RAM to store them. Once we had the instance running, we ran two types of tests to evaluate the authorship model's performance.

Overall Accuracy Test

Our first test was on the authorship model's overall accuracy when trained on inputs from all of the datasets at once. Due to limited resources and the fact that creating 10,977 embeddings required 591 GB of RAM, we limited the number of examples we used to a maximum of 2000 per dataset. After running this test, the authorship model produced an overall accuracy of 81.86% on the test set. This accuracy shows that the model is able to classify many different types of authors correctly, but the misclassified 18% implies that this performance comes with caveats. Many of the examples that the model trained on were in the mixed category, so it is likely that a large number of misclassified examples exist because the model is unable to fully comprehend the difference between fantasy and reality like a human can. Specific datasets that do not work well with the classifier could also bring the overall accuracy down. Our second test examines this possibility.

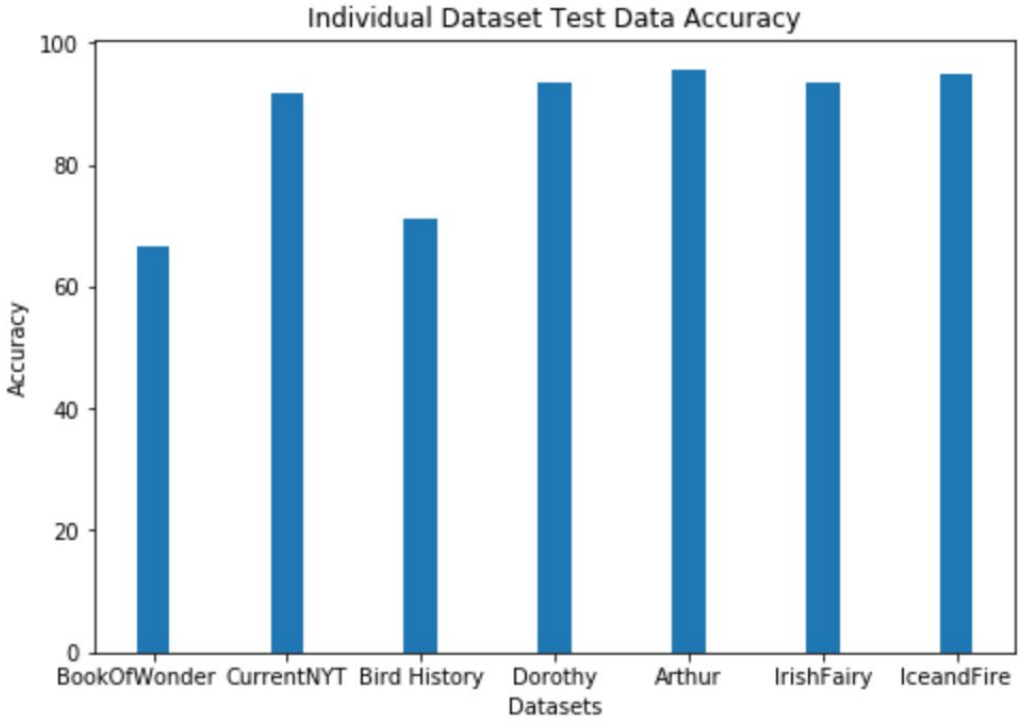

Figure 1

Individual Accuracy Test

The second test involved taking all of the sentences from each individual dataset and determining the accuracy of the authorship model when classifying between only that author and a mixture between realistic and fantasy data. The results of these tests are plotted in Figure 1. Through the results of these tests, we discovered that the Book of Wonder and Bird History datasets exhibited poor performance on the task. Although this result can be partially attributed to the lower amount of training data that these datasets had, the lower accuracy achieved by the Book of Wonder dataset compared to the Bird History dataset is important to address. Despite having more available training examples, the Book of Wonder dataset achieved lower accuracy on the task. This indicates that the dataset is similar enough to realistic datasets that the authorship model can not distinguish between them, but the difference in content still significantly affects the end result of the classifier. Consequently, the low performance on the authorship task is an indicator that the style of a book in a fantasy domain can negatively affect performance of downstream tasks. In contrast, the authorship classifier achieved very high accuracy on the World of Ice and Fire data despite the similarity of that writing to realistic data. This shows that the influence that a difference in style has is hidden away in the embeddings and is not necessarily measurable. The presence of these hidden influences means that tasks trained on data from fantasy domains may exhibit unreliable behavior.

Conclusion

Throughout the experiments there were two noteworthy trends in how BERT interacted with the raw fantasy data. The first trend pertains to the effect that style has on the BERT embeddings. Even though the Arthur dataset's style did not improve performance on the Next Sentence Prediction Task, the BERT model was able to pick up on it and produce high accuracy on discriminative tasks. Additionally, this behavior was not seen to the same extent on the Irish Fairy Tales dataset, which also contains many uncommon terms in it but does not use as prominent of an old English style. Consequently, these results show that a significantly different style is identifiable with a BERT model. The second trend pertains to the effect that mixing styles has on the BERT data. As was noted earlier in the analysis of the authorship classifier, mixed styles exert an influence on the embedding that is hidden away as if the data is from a realistic domain. This means that models can not easily distinguish between where the realistic style and the fantasy elements begin.

Without an inherent sparsity between fantasy data and realistic data, it is likely that these influences will propagate into downstream tasks and decrease performance. Similar to what was seen in the Authorship test on the Book of Wonder dataset, these slight differences could significantly impact the accuracy of downstream tasks. Ultimately, our results show that not all fantasy data is created equal. Different styles

of fantasy will produce different effects on fine-tuning tasks, and BERT embeddings do not provide full interpretability on how these differences manifest themselves. Further analysis of these effects would need to be done through experiments using downstream tasks that use labeled data.

References

Research Papers

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 1.Jacob Devlin2. Ming-Wei Chang3. Kenton Lee 4. Kristina Toutanova. https://arxiv.org/pdf/1810.04805.pdf

[2]Multi-classifier system for authorship verification task using word embeddings. 1.Nacer Eddine Benzebouchi 2.Nabiha Azizi 3. Monther Aldwairi 4.Nadir Farah https://ieeexplore.ieee.org/abstract/document/8374391

Other Resources

- https://gluebenchmark.com/leaderboard
- https://keras.io/layers/recurrent/
- https://www.gutenberg.org/wiki/Main_Page
- https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
- https://github.com/jalammar/jalammar.github.io/blob/master/notebooks/bert/A_Visual_Notebook_to_Using_BERT_for_the_First_Time.ipynb