Distinguishing 19th Century British Novels by Women Authors Using Natural Language Processing

Phoebe M. Xu*
North Hollywood High School Highly Gifted Magnet
North Hollywood, California

Abstract

This paper utilized the BERT model and binary logistic regression to distinguish books written by 19th-century British women, specifically exploring AI's ability to determine author differences and keywords in each book. Two books each by Jane Austen, Mary Shelley, and Mary Brunton were divided into uniformly sized sections to train and test the BERT model. Its task was to analyze the author-labeled training set, and then assign author labels to the separate testing set. The results showed that the model achieved 84.44% accuracy. A z-test yielded a z-score of 35.63 and a negligibly small p-value approaching 0. Binary logistic regression was then utilized to pinpoint the most distinctive words from each book, helping to understand the differences between the books.

Keywords: Authorship, Jane Austen, Mary Shelley, Mary Brunton, Natural Language Processing, BERT model, Binary Logistic Regression

1 Introduction

In today's world, artificial intelligence (AI) has become a technology with many areas of ongoing growth and innovation. Many of its mechanisms are still unknown and developable since its systems can learn and process data alone. Additionally, they are important to study due to their widespread applications in society, described by researchers from the Hague Center for Strategic Studies, "Usage ranges from fairly quotidian comparisons of inter-human skill or markers of authority, to extremely theoretical definitions of intelligence..." (De Spiegeleire et al. 2017). One specific area of such study is natural language processing, and this paper will specifically look at its applications to 19th-century British female authors.

*phoebexu2025@gmail.com

1.1 Natural Language Processing

Natural language processing focuses on the specific ability of AI to process words and make determinations based on text. It utilizes pattern recognition and sentence parsing, with different models executing this differently (Coughlin, 1990). Recent innovations have developed new AI pattern recognition techniques profoundly different from human approaches, opening doors to test for unfamiliar natural language processing abilities (Holzinger et al., 2023). Through the careful analysis of the effectiveness of natural language processors, a clearer understanding of their learning techniques can then be developed into real-world applications (Marcus, 1995). An example was the study of max-margin structures, a type of natural language processor that could determine parts of speech with high precision; and statistical machine translation, which could directly map text from one language to another. These two fields evolved to become the basis of many translation systems used around the world today (Ni et al., 2010). Other work with natural language processing has even intersected with other aspects of technology, such as sound and audio processing, to create speech-to-text technology (Hirschberg & Manning, 2015). The specific natural language processor in this project was a pre-trained language model called the Bidirectional Encoder Representations from Transformers, or BERT for short. It specializes in learning text patterns associated with certain labels and then labeling other text based on what it was trained on (Bao et al., 2021). This type of categorization and labeling is called sentiment analysis, which in this project, comes in the form of determining authorship.

The applications of sentiment analysis can vary, but they all center around one specific method of analyzing text: the classification of text into three categories, or sentiments. These three categories are most commonly labeled as "positive", "neutral", and "negative", though specific pre-trained models can also have other categories specific to the types of text it was trained to classify. This kind of analysis allows for a greater understanding of polarity in writing and the consumed information, although it also leads to discrepancies when looking at biased information (Pang & Lee 2004). An example of research where sentiment analysis played a particularly effective role in determining trends was at the North Carolina AT State University, where researchers wanted to determine whether AI could classify Amazon product reviews as positive, neutral, and negative, based on context and emotions associated with words in the reviews (Fang & Zhan, 2015).

1.2 The Intersection of Natural Language Processing and 19th Century Female British Authors

Previous research analyzing authorship with sentiment analysis has been conducted in many different contexts. One example was done by professors in the United Kingdom and the Netherlands who were trying to see if BERT models could determine the authorship of different late-19th-century novels (Silva et al., 2023). They utilized a model trained to become accustomed to the writing styles of the different authors and then used the model to predict who wrote other manuscripts. Their findings had largely come out positive, although there were variations based on external factors. A significant impact on their study by such external factors was their sample size–something that other studies had also experienced. One of these included a project by professors at the University of Turku to predict the

publication year of 18th-century writing. Throughout the study, not only did they analyze the ability of the BERT model, but also the optimal sample size usage to train the BERT model for accurate results (Rastas et al., 2022). Other researchers believe that BERT models and natural language processing cannot make such inferences without significant issues–especially when applied to broader societal contexts. Professors at Rush University found that biases in coding their BERT models had resulted in "disadvantage [to] some groups or populations over others—often those already disproportionately marginalized" (Thompson et al., 2021). Thus, these issues must be addressed and worked around when coding towards a specific goal.

1.3 Gaps in Current Research

The gap in current research that this manuscript addressed had already been explored but as separate projects and never altogether. Although late 19th century books had been analyzed with a BERT model for authorship before, they had yet to be related in author demographic. This paper tested whether natural language processing could determine the authorship of books written by three close in proximity yet distinct female authors of early 19th century Britain: Jane Austen, Mary Shelley, and Mary Brunton. The books chosen from these authors were all published between the years 1811 and 1823, showing a narrow range in timing that could potentially challenge the BERT model. By addressing this gap in research, we can better understand the abilities of AI and the challenges it could overcome when faced with socially complex novels of proximity, and broaden the possibilities of future applications of said AI. Thus, the question that this paper aimed to answer was: Can artificial intelligence and statistical modeling accurately differentiate between the authorship of books written by early 19th-century British female authors?

2 BERT Model Methods

This manuscript utilized books written by three British female authors from the early 19th century to train and test a BERT model. The books were *Sense and Sensibility* and *Pride and Prejudice* by Jane Austen, *Frankenstein*, and *Valperga* by Mary Shelley, and *Self Control* and *Discipline* by Mary Brunton. A Python BERT model was coded to analyze the novels, which was trained on sections from the beginning of each book, and then tested on sections from later in the novels. The BERT model's success rate was then analyzed with a z-test for proportions, which determined whether or not the model was more successful than random guessing. A 95% confidence interval was also constructed to approximate the model success rate, inferred from sample data.

Essentially, a BERT model works through a transformer by utilizing a self-attention mechanism to learn information from a piece of text in both directions–as outlined in its name, "bidirectional". A transformer is essentially a deep-learning mechanism which converts text into numerical "tokens" through the following method. These tokens are then given in an input sequence $x = (x_1, x_2, ..., x_n)$ where x_i represents the embedding of each input, the mechanism computes attention scores α_{ij} for the inputs in two positions i and j defined as:

$$\alpha_{ij} = \frac{exp(e_{ij}}{\sum_{k=1}^{n} exp(e_{ik})}$$

where $e_{ij} = \frac{(x_i W_Q) * (x_j W_K)^T}{\sqrt{d_k}}$. Here, W_Q and W_K represent projection matrices learned from the training input sequence that transform the embedding into query vectors and key vectors. The dimensions of these key vectors are represented by d_k. Thus, α_{ij} can be used to calculate a weighted sum of the vectors $v_j = x_j W_V$ as represented below:

$$z_i = \sum_{j=1}^{n} a_{ij} v_j$$

where W_V is a learned projection matrix for the vectors. The subsequent output vectors from this, z_i, are then processed in a feed-forward structured neural network. Through this, BERT develops a bidirectional approach that allows it to analyze the left and right context of words. The overall understanding and quality of the text is thus improved.

The BERT model trained on uniform-length text strings in bulk, where each string was labeled by an author's name. This allowed it to learn and distinguish between specific patterns in different authors' works. These patterns then allowed the model to determine the authorship of other unlabeled text strings, all of which have uniform word counts to minimize coded biases. The model was coded specifically for this paper, but a separate code had to be designed to format the text strings in both the training and testing sets. This program deleted all the punctuation and capital letters from the books and separated them into strings of equal length for maximum training and testing efficiency. Specifically, professors at the University of Turku previously discovered that sections of 512 words in length were the most effective at training a BERT model, which was applied in this paper (Rastas et al., 2022). As for testing, each string was 128 words long to yield data that could be sufficiently analyzed later on.

2.1 Coding the BERT Model

This paper utilized a BERT model that was created in Python, specifically formatted so that its inputted data for both training and testing would come from comma-separated values. There were two separate parts to this model, specifically a training code and a testing code. The training code took a CSV file with data entries of the same length and categories in the second column to learn patterns within each categorization. Then, it would associate those patterns with the categories. In this situation, the categories were the names of the authors. The testing code would then also take input in the form of a CSV file, except this time its uniform data entries did not include categorizations. The BERT model would have to analyze each entry, and based on patterns observed before, assign a category to each entry. Please refer to the Appendix to see the training code of the BERT model, and the testing code of the model.

These models need to be rerun every time, meaning that they do not store information from past experiments. Thus, this model was run a few times, not as a part of the experiment method, to test out the code and make any adjustments necessary, but not in a way that affected its learning abilities.

2.2 Training the BERT Model

As mentioned earlier, the books were divided into sections of 512 words each to determine the effectiveness and accuracy of the model under different training circumstances. This was done using a separate code in Python, which took out all of the punctuation, capitalization, and line breaks to make strings of 512 words long from the books. The code for this can be referred to below in the Appendix. Once divided, the novels were stored as comma-separated values to be read by the model, as described earlier to be the training set. Since the books are of different lengths, to standardize the datasets of each author, 100 sections of 512 words were utilized from each novel.

2.3 Testing the BERT Model

After training the model, BERT then looked at the testing set to determine the authorship of the data utilizing prior observations. Since only the first part of the novels was used to train the model, the latter portions of the novels were divided into uniform text strings and tested in the model. This also needed to be standardized, which was why 180 sections were taken from each book, with one section being 128 words long. This was established as each data entry would be effectively smaller than the training data entries, yet still distinguishable from each other to be categorized with maximum efficiency. The importance of this step was to see whether the model could accurately recognize the authorship of the text it was trained on as well as new text from the same author.

2.4 Analyzing the Data

From this categorical data, the proportions of success in this specific sample for determining BERT effectiveness were statistically evaluated. A z-test for proportions was conducted utilizing the success rate of this sample compared to the expected proportion of success of an ineffective model, which is 33.33%. This provided perspective on the ability of the model to determine the authorship of the text samples it analyzed, and how likely it was to be more effective than random guessing. In addition to this test, a 95% confidence interval was also constructed with the sample to determine an estimated interval of what the model success rate is, with inference to the model in general. This provided not only an understanding of whether the model was effective but also measures the degree of effectiveness of the model.

3 BERT Model Results

This model was run according to the methods above, yielding data that potentially suggested many interesting implications. The number of successes in the raw data is shown below in the following chart:

Author:	Jane Austen		Mary Shelley		Mary Brunton	
Book:	Sense and Sensibility	Pride and Prejudice	Frankenstein	Valperga	Self Control	Discipline
# Correct:	171	171	158	158	127	127
Author Total:	342		316		254	

From these results, it is clear that the model produced a majority of correct responses. The most successful out of these had been Jane Austen's novels, which had 95.00% of its entries identified correctly. For both Mary Shelley, who had 87.78% of entries correct, and Mary Brunton, who had 70.56% of entries correct, their novels had lower success rates in the BERT model. Upon further observation of the individual data points and errors, many of the Mary Shelley and Mary Brunton errors were identified as each other. This suggests possible implications surrounding the patterns that BERT analyzes, perhaps relating to the specific books. In addition, all 18 Jane Austen errors were attributed to Mary Brunton, suggesting possible patterns between the writing of these two authors as well. However, the most prominent anomaly that occurred in all of the data was the identical success rates for each of the books by the same author. This was observed among all three authors, where both of the books from each of them yielded the same number of correct AI determinations.

4 Discussion and Conclusion

To analyze this data, a z-test for proportions was conducted to determine whether or not the probability of success utilizing the BERT model was greater than the probability of random guessing alone. Random guessing would have yielded an expected value of 33.33% successes, but the model had an overall correct entry proportion of 84.44% and individual successes of 95.00%, 87.78%, and 70.56% each. Conducting the z-test with the overall sample proportion yields a z-score of 35.63, yielding a resulting p-value negligibly larger than 0. The interpretation of such a p-value is that it is smaller than 0.05, which signifies the statistical significance of the success rate of the model. The null hypothesis that the model was just as effective as random guessing can be rejected, and the alternative hypothesis that the model was more effective than random guessing can be accepted. This proves that the AI BERT model was clearly effective in determining the authorship of 19th-century British Female-written books. To estimate a specific range containing the rate of success of the model itself, a 95% confidence interval was created utilizing the data. With the 1080 data entries of either successes or failures and the rate of success of 84.44%, the confidence interval was constructed to be approximately 0.8444 ± 0.0230. In other words, the general success rate of the BERT model should be between 82.14% and 86.56%. Such knowledge helps to expand our knowledge on the abilities of BERT outside of just comparison to random guessing, but also answering the question of how successful the model itself was and whether or not it can be categorized as accurate as asked by the research question. Utilizing the scale described in the methods section, it can be interpreted that the model was accurate. The entire confidence interval falls within the 81-100% level, leading to this conclusion.

In addition to quantitative data analysis, which was used to make inferences and interpretations of the data provided, qualitative observations were also made to explore what could be researched in this field in the future. One of the most prominent of these observations was the mistaking of Mary Shelley and Mary Brunton novels, which occurred as one of the most common errors in the entire data set–both in Shelley's data entries and Brunton's data entries. A possible cause for this trend could have been the sharing of proper nouns across their books, such as names of people and places. As the BERT model analyzed patterns, the presence of similarities could have confused its recognition abilities, but proving this fact is an area of future research that could be done.

In terms of other patterns that BERT may have been affected by, one factor that should be considered is genre similarities. While Mary Shelley's novels consisted of gothic horror and historical fiction, Jane Austen and Mary Brunton both wrote Regency-era novels that discussed topics such as familial standing, marriage, and social norms. This may have been related to the anomaly that all of the errors in the Jane Austen novels were identified and categorized as Mary Brunton novels. Despite this trend, Jane Austen's novels were still the highest performing, suggesting a distinctive nature to her writing that sets her apart from other authors despite genre similarities. For the future, though, looking into the relationships between AI and genre could also be useful in technological development regarding BERT models.

5 Binary Logistic Regression

As for the statistical test in this paper, a logistic regression model in R was used to determine the most frequent words of each author when compared to another author. The same training sets of sections of 512 words each from the BERT model were used for this test as well, but divided into three trials comparing two authors to each other at a time. This consists of the following three groups: Jane Austen and Mary Shelley, Mary Shelley and Mary Brunton, and Jane Austen and Mary Brunton. After determining the 50 most frequent words in the training set (minus "stop words", which include words less than 4 letters in length and other frequent parts of speech such as "that"), the code then determined whether or not the words were statistically significant to one author or not. "Statistically Significant Words" meant that if a certain word was chosen randomly, the probability that the word came from a book of one certain author is very high, and thus, statistical significance. The following is the theoretical framework for the binary logistic regression model in this paper.

5.1 Binary Logistic Regression Theoretical Framework

In binary logistic regression, the response variable Y is assumed to have a Bernoulli distribution with the parameter π. Thus, Y can assume values 0 or 1, and the probability of $Y = 1$ is $\pi = \mathbb{P}(Y = 1)$. Then, π is modeled as a function of predictor variables $x_1, \ldots, x_k$ via the logistic function $f(x) = e^x/(1 + e^x),\ -\infty < x < \infty$. It can be written as

$$\pi = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\right)}$$

where $\beta_0, \ldots, \beta_k$ are the regression coefficients estimated from the data using the method of maximum likelihood. Define the odds in favor of $Y = 1$ as

$$\text{odds} = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} = \frac{\pi}{1-\pi}.$$

The fitted model can be written in terms of the odds as follows

$$\frac{\hat{\pi}}{1-\hat{\pi}} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k\right).$$

In the context of this paper, all predictors are indicator functions of whether particular words were used, and thus, they are all 0-1 variables. The ratio of the estimated odds when the predictor variable, x_1, is equal to 1 vs. when it is equal to 0 can be found as

$$\frac{\widehat{\text{odds}}_{x_1=1}}{\widehat{\text{odds}}_{x_1=0}} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 (1) + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k\right)}{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 (0) + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k\right)} = \exp(\hat{\beta}_1).$$

If the estimated beta coefficient $\hat{\beta}_1$ is positive, the odds in favor of $Y = 1$ are larger when $x_1 = 1$ than those when $x_1 = 0$.

In the analysis of this paper, an author (Author 1) ($Y = 1$) is regressed against another author (Author 2) ($Y = 0$) on words used in their books. Words with positive estimated regression coefficients are more likely to belong to Author 1, while words with negative estimated beta coefficients are more likely to be attributed to Author 2. The words with statistically insignificant regression coefficients (p-values> 0.05) can be encountered in works by both authors and are not useful in distinguishing the authorship.

6 Binary Logistic Regression Results and Discussion

The binary logistic regression yielded an analysis of the words that distinguish between pairs of authors. See the charts in the Appendix for lists of these significant words. One of the main anomalies in the data was that regarding names and proper nouns, which was almost always purely unique to one book alone. By having such words commonly appear, the model cannot generate odds and statistical significance predictions since it would be dividing by zero. Thus, in the charts above, the names and proper nouns are all crossed out. However, even though a few of the names were shared, they were also crossed out for uniformity in the study.

Overall, there was a large overlap between words from the same author even when compared to two different authors, except for a few cases. Anomalies such as the lack of significant words from Jane Austen when compared to Mary Brunton can possibly be explained by genre similarities, where both of their works cover similar topics and constructs. Mathematically, there is a strong positive correlation between the presence of significant words, as found through binary logistic regression, and the success of the BERT model in determining authorship in a given section. However, what was still the most apparent and expected result was that words pertaining to the stories of each novel would be the most prevalent. For

example, when comparing Mary Brunton and Jane Austen, many words were not significant to Jane Austen due to the fact that their genres were similar. However, since Jane Austen wrote novels about sisters, such as *Pride and Prejudice*, "sister" was then a significant word.

7 Implications, Limitations, and Future Considerations

The applications of analyzing BERT models can have many potential implications for future usage and consideration in the world today. The most immediate implication of this paper specifically regards the research gap this paper addressed, which was the usage of books written by authors of similar demographics. Since all six books were published within 13 years of each other, the BERT model was challenged to distinguish between novels written under similar social circumstances. This is only further emphasized by the fact that all three of the authors were women. In the early 19th century, there would have been many more expectations, norms, and trends women were obligated to follow that could have impacted the type of writing present in their novels. Thus, by using such a sample of writing in this manuscript, the BERT model's limits and abilities were able to be understood better.

Another implication of this paper includes the role of AI in the writing community. Currently, artificial intelligence is often seen as a hindrance to writers due to its ability to generate text mimicking humans, which is harmful to careers and creatives. However, through technology such as BERT models and natural language processing, AI becomes a tool that can be developed to help our understanding of literature and writing. In the future, these applications could include abilities such as detecting plagiarism utilizing writing styles, as well as determining the authorship of anonymously written and published pieces throughout history. Through these applications, artificial intelligence is no longer limited to the scope of current technology, but can also be extended to broader areas of literary research for the future. Many examples of such research also relate to the limitations and future considerations regarding this manuscript.

To discuss future considerations for research regarding BERT models and natural language processing, the limitations of this paper can be analyzed and extended. One of the main downfalls of this paper was its inability to analyze the specific patterns and trends of the BERT model itself. This manuscript was only able to determine the accuracy of the BERT model, but future research can be very valuable in understanding how the model analyzes patterns in text. Through these unknown patterns, it is also possible that the training data utilized in this paper could have been inadvertently biased. Many efforts were made to minimize biases, such as standardizing the sample size and text lengths when training and testing the model, as well as other aspects of the code that were designed to minimize external impacts. However, in the end, this limitation to the methods cannot be entirely avoided, but future work analyzing such can greatly benefit the technology. For example, some of the specific potential impacting factors could have included overlap in proper nouns, which occured in Mary Shelley and Mary Brunton's novels, or varying book lengths. Such situations could have easily influenced patterns in the writing, which would in turn affect the categorization ability of the BERT model due to their obvious impact on the way the model is trained. Thus, extending and building upon BERT model research can largely be beneficial to our understanding of writing, modern and historical.

Acknowledgement

I would like to thank Dr. Olga Korosteleva from the California State University of Long Beach for her advice on this paper, including guidance with the BERT model and the Binary Logistic Regression theory. Her advice and mentorship have helped momentously for this project and it is greatly appreciated.

References

Bao, H., He, K., Yin, X., Li, X. (2021). BERT-based meta-learning approach with looking back for sentiment analysis of literary book reviews. Natural Language Processing and Chinese Computing, 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II (235-247). DOI: 10.1007/978-3-030-88483-3_18

Coughlin, J. (1990). Perspectives on natural language processing. The French Review, 64(1), 172–179. http://www.jstor.org/stable/395699

De Spiegeleire, S., Maas, M., Sweijs, T. (2017). What is artificial intelligence? In artificial intelligence and the future of defense: Strategic implications for small- and medium-sized force providers. Hague Centre for Strategic Studies, 25–42. http://www.jstor.org/stable/resrep12564.7

Fang, X., Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data 2(5). https://doi.org/10.1186/s40537-015-0015-2

Fields, P. J., Bassist, L. W., Roper, M. R. (2017). Characters in 19th-century novels display distinctive voices as seen by stylometric analysis. Brigham Young University. https://dh2017.adho.org/abstracts/494/494.pdf

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A corpus-stylistic analysis. International Journal of Corpus Linguistics, 14(4), 492–523. DOI: 10.1075/ijcl.14.4.03fis

Hirschberg, J., Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261–266. http://www.jstor.org/stable/24748572

Holzinger, A., Saranti, A., Angerschmid, A., Finzel, B., Schmid, U., Mueller, H. (2023). Toward human-level concept learning: Pattern benchmarking for AI algorithms. Patterns (New York, N.Y.), 4(8), 100788. https://doi.org/10.1016/j.patter.2023.100788

Marcus, M. (1995). New trends in natural language processing: Statistical natural language processing. Proceedings of the National Academy of Sciences of the United States of America, 92(22), 10052–10059. http://www.jstor.org/stable/2368613

Ni, Y., Saunders, C., Szedmak, S., Niranjan, M. (2010). The application of structured learn-

ing in natural language processing. Machine Translation, 24(2), 71–85. http://www.jstor.org-/stable/40926416

Pang, B., Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04) (271–278). ACL. DOI: 10.3115/1218955.1218990

Patowary, U. (2023, August 18). Artificial intelligence and Mary Shelley's Frankenstein: A comparative analysis of creation, morality and responsibility. Integrated Journal for Research in Arts and Humanities, 3(4), 121–127. https://ssrn.com/abstract=4544608

Rastas, I., Ryan, Y. C., Tiihonen, I., Qaraei, M., Repo, L., Babbar, R., Mäkelä, E., Tolonen, M., Ginter, F. (2022). Explainable publication year prediction of eighteenth-century texts with the BERT model. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, 68–77. Dublin, Ireland: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.lchange-1.7

Silva, K., Can, B., Blain, F., Sarwar, R., Ugolini, L., Mitkov, R. (2023). Authorship attribution of late 19th century novels using GAN-BERT. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), 310–320. Toronto, Canada: Association for Computational Linguistics.

Thompson, H. M., Sharma, B., Bhalla, S., Boley, R., McCluskey, C., Dligach, D., Churpek, M. M., Karnik, N. S., Afshar, M. (2021). Bias and fairness assessment of a natural language processing opioid misuse classifier: Detection and mitigation of electronic health record data disadvantages across racial subgroups. Journal of the American Medical Informatics Association, 28(11), 2393–2403. https://doi.org/10.1093/jamia/ocab148

Appendix

The following code was used to construct the training portion of the BERT model.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from simpletransformers.classification import ClassificationModel

# Read data from CSV file
data = pd.read_csv("training_set.csv", encoding='ISO-8859-1')

# Map authors to numerical labels
def map_author_to_label(author):
    if author == 'JaneAusten':
        return 0
    elif author == 'MaryShelley':
        return 1
    else:
        return 2

data['label'] = data['author'].apply(map_author_to_label)

# Split the data into training and testing sets
train_data, test_data = train_test_split(data, test_size=0.2, random_state=42)

# Create a TransformerModel
model = ClassificationModel(
    'bert',
    'bert-base-cased',
    num_labels=3,
    args={'reprocess_input_data': True, 'overwrite_output_dir': True},
    use_cuda=False
)

# Train the model
model.train_model(train_data[['text', 'label']])
```

The following code was used to construct the testing portion of the BERT model.

```
import numpy
import pandas as pd

#using the trained model to classify user-defined sentences
def classify(statement):
        result = model.predict([statement])
        pred_class = numpy.where(result[1][0] == numpy.amax(result[1][0]))
        pred_class = int(pred_class[0])
        sentiment_dict = {0:'Jane Austen',1:'Mary Shelley',2:'Mary Brunton'}
        print(sentiment_dict[pred_class])
        return

stuff = pd.read_csv("testing_set.csv",encoding='ISO-8859-1')

for index, row in stuff.iterrows():
    statement = row['text']
    classify(statement)
```

The following table depicts words significant to Mary Shelley and Jane Austen from among the top 50 most frequent words in the training set.

	Mary Shelley	**Jane Austen**
p<0.001	even	lady
	eyes	miss
	father	much
	felt	nothing
	first	sister
	friend	think
	heart	though
	life	
	love	
	made	
	many	
	upon	
0.001<p<0.01	ever	know
	time	mother
0.01<p<0.05	might	said
	shall	without
	thought	

The following table depicts words significant to Mary Brunton and Jane Austen from among the top 50 most frequent words in the training set.

	Mary Brunton	**Jane Austen**
p<0.001	cried	sister
	even	
	father	
	first	
	friend	
	indeed	
	lady	
	like	
	little	
	love	
	made	
	make	
	might	
	miss	
	never	
	pleasure	
	said	
	thought	
	upon	
	without	
0.001<p<0.01	ever	
	however	
	though	
	time	
0.01<p<0.05	long	

The following table depicts words significant to Mary Brunton and Mary Brunton from among the top 50 most frequent words in the training set.

	Mary Brunton	Mary Shelley
p<0.001	captain	even
	lady	ever
	miss	every
	though	eyes
		father
		felt
		first
		found
		friend
		good
		heart
		life
		like
		love
		made
		many
		might
		mind
		must
		often
		passed
		seemed
		shall
		soon
		still
		time
		towards
		well
		whose
0.001<p<0.01	lord	indeed
		never
		pleasure
		thought
		upon
0.01<p<0.05		little
		much