Classification of Time Period and Author Age in Fiction

ABIGAIL HODGE, Northeastern University
SAMANTHA PRICE, Northeastern University
ZIJIN HUANG, Northeastern University

ACM Reference Format:
Abigail Hodge, Samantha Price, and Zijin Huang. 2019. Classification of Time Period and Author Age in Fiction. 1, 1 (April 2019), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Our task is to classify a given English novel into one of the following 7 time period: 1751-1800, 1801-1820, 1821-1840, 1841-1860, 1861-1880, 1881-1900, 1901-1920; and one of the following 5 author age category (at the time of writing the novel): 18-24, 25-34, 35-49, 50-64, 65+.

System input: a .txt file containing the original text of the English novel.

System output: a number from 0 to 6 representing the 7 time period categories, and a number from 0 to 4 representing the 5 author age categories.

For example, given a text file containing the romantic novel Pride and Prejudice by Jane Austen, the system should output 2 for time period (1813 is in the 1821-1840 time period category), and output 2 for age category (born 1775, age 38 when writing the book, in the 35-49 age category).

2 RELATED WORK

This section discusses previous work associated with author age identification and time period identification.

2.1 Nguyen et. al. (2011) & Rangel and Rosso (2013)

Nguyen et. al. (2011) [2] examined three corpora — blog posts, telephone conversations and forum discussions — to create a linear regression model for author's age. The main concern of this paper was creating a generalized model that could be applied to many different styles of writing (although notably not fiction) as well as modeling age as a continuous problem rather than a category problem. The model utilizes both grammatical and content features, including POS, slang usage, and sentiment analysis. We are interested in this paper primarily for its treatment of age as a spectrum, which we might utilize in our own model. Rangel and Rosso (2013) [5] also examined author's age, although they used a SVM model and different features, such as punctuation and emoticons, which might be worth considering.

2.2 Garcia-Fernandez et. al. (2011)

Garcia-Fernandez et. al. (2011) [1] created an SVM model to determine year of publication for French newspapers. This was a significantly more specific scope than we are interested in, as they determined exact year and we are only interested in decade. However, they utilized named entities, neologisms, and anarchism in an intriguing way, which we would like to replicate or improve upon in our model (more in Section 4).

2.3 Khosmood and Levinson (2008)

Khosmood and Levinson (2008) [4] focused on classification and transformation of writing style in a variety of documents. For classification they defined a list of style markers (such as frequency of pronouns) based on their training corpora and identified these same markers in a test document. Transformation involved a comparison of styles between a source document and a target document. The style of the source document was gradually modified until it matched (or nearly matched) the style of the target document. Although our models will not involve a direct analysis of style, the concept of style markers that Khosmood and Levinson described could prove useful in our prediction of author age and range of publication years.

2.4 Gamon (2004)

Gamon (2004) [3] also discussed style classification like Khosmood and Levinson (2008) [4], but for the purposes of authorship identification rather than style transformation. He utilized both "shallow" and "deep" linguistic features to accurately determine the author of a given document; each document included lines selected from one of the Bronte sisters' works. Again, style classification will play a role in our task, albeit for a different outcome than Gamon's work.

2.5 Our project

Our project is unique in that it addresses both age category and time period category comprehensively, indicates which factors lead to a work's classification, and analyzes novels instead of newspapers, blog posts, or social media. The more formal writing style and careful editing of a novel had lead to alternative feature vectors and a different form of analysis.

3 DATASET

The dataset we utilized was a subset of novels (in text file format) from Project Gutenberg. Each novel in the dataset fulfills a set of defined criteria: originally published in English between 1750-1900, with a single author who has a known birth year. The dataset

Authors' addresses: Abigail Hodge, abigailhodge98@gmail.com, Northeastern University; Samantha Price, price.sam@husky.neu.edu, Northeastern University; Zijin Huang, elizazijinhuang@gmail.com, Northeastern University.

XXXX-XXXX/2019/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

contained a total of 380 books. The time range of 1750-1900 was defined as such because books published within that time frame are currently available in the public domain (and thus likely to be found in Project Gutenberg's collection).

4 METHODOLOGY

4.1 Preprocessing

To extract the subset of novels, we first generated a local copy of the Standardized Project Gutenberg Corpus (SPGC) containing 59221 books in .txt format and a metadata.csv documenting related information (id, title, author, author year of birth author year of death, language, downloads, subjects type).

Then we removing the header, footer and other separation symbols added by project Gutenberg, and sorted the books by descending number of downloads (rank by popularity).

Lastly, we used Wikipedia API to extract the publication year for each text; computed age at the time of publication, and classify each book into a time period category and age category. In the process, we removed all text that are not novel, not in English, does not have a specific publication year from Wikipedia, published outside the range 1750-1900, or does not have an indisputable author birth year.

4.2 Feature Extraction and Selection

Our strategy for feature selection was to initially try any features that we believed would be helpful, then remove unnecessary noise using a random forest model. To that end, we selected a wide variety of content, style, semantic, and syntactic features for both age and time period.

Related works for both age and time period determination indicated that unigram counts and POS tag counts would be useful. The top 500 unigrams were extracted from our test corpus and used as a vocabulary. The tf-idf for each word in this vocabulary was then calculated for each document. We utilized the Stanford POS tagger to get POS counts for each document.

To examine the semantics, we trained a Word2Vec model on our training corpus. We considered using the GoogleNews pretrained embeddings, but we believed that better results could be achieved using a more domain-specific model, as language use is very different between novels and newspapers. Once the model was trained, we averaged together the embeddings of each word in each document. We also used Word2Vec to determine the average sentence similarity in each document.

We suspected that topic classification might also be useful for both age and time period classification. Younger writers might discuss different ideas than their older counterparts, and themes shift over the ages. We built an LDA model for our training corpus, then determined the topic makeup of each document.

Our final feature for both age and time period was vocabulary size, which we determined by comparing the number of unique words in each document with the document's total number of words.

We added a few additional features for time period classification. The first was named entity counts. We ran the Stanford NER on our training corpus, then similarly to our process for unigram counting, we created a vocabulary of the top 100 named entities, and found the tf-idf values for said vocabulary in each document. Finally, we found the Flesch readability score for each document.

4.3 Model

4.3.1 Linear Regression. When we began our preliminary research into age and time period classification, we determined that a simple linear regression model would be sufficient, based upon previous research (CITATION HERE). However, this model proved to be extremely ineffective in predicting exact age and time period of any given novel. Thus we determined that we would predict age range and time period classes instead of exact values. With this alteration and the large number of features that needed to be analyzed, we decided that a feedforward neural network would be more appropriate.

4.3.2 Multilayer Perceptron. We created a multilayer perceptron with a single hidden layer to classify novels as one of 5 different age ranges and one of 6 different time periods (defined in Section 1). We utilized 5-fold cross validation on a training set containing novels by various authors (extracted automatically from Project Gutenberg as described in Section 4.1) With a dataset size of approximately 64 novels, it was determined through cross validation that the optimal hyperparameters for the age classifier included a relu activation function and a Stochastic Gradient Descent optimizer. For the time period classifier, a tanh activation function and a limited memory BFGS optimizer were preferred. The metrics produced from the age classifier and time period classifier can be found in Section 6.1.1.

One of our major goals for this project was to determine which specific features were most influential in selecting an age range and a time period for any given novel. With our choice of a multilayer perceptron, it was difficult to extract this information. Thus, we ultimately chose to select a third and final model for classification, one that would allow us to retrieve feature importance information: a Random Forest classifier.

4.3.3 Random Forest. The key reason we chose to utilize a Random Forest classifier as our final model was the easy accessibility to feature importance that it provides. With this access, we were able to isolate the features that had the largest impact in classification, as well as drop the features that seemed to have very little impact. The process in training the model was extremely similar to that for the multilayer perceptron; the only major difference was the size of the initial dataset (380 novels). After splitting the dataset into training and test sets, 5-fold cross validation was performed on the training set, with a focus on optimizing the number of trees in the random forest, the maximum depth of each tree, and the maximum number of features to be considered when looking for the best split. (FILL IN OPTIMIZED VALUES HERE). Metrics for the age and time period Random Tree classifiers can be found in Section 6.1.2.

5 EXPERIMENTS

5.1 Dataset and Metric

For both the Multilayer Perceptron and the Random Forest classifier, the baseline metrics were determined by a "dummy" classifier, specifically provided by Scikit-Learn to test against other models.

The metrics utilized to evaluate both models were accuracy, precision, recall, and the F1 Score. Accuracy was chosen as a simple and straightforward determination of how capable the models were in classifying correctly, and the other aforementioned measures were included to provide a comprehensive evaluation, as accuracy can be easily skewed.

5.1.1 Multilayer Perceptron. As previously mentioned, the dataset for our experiments with the multilayer perceptron was comprised of 64 novels from 22 different authors. The dataset was split into 51 books for training (80%) and 13 novels for testing (20%). The baseline accuracy for age (provided by the dummy classifier) was 0%. The baseline for the other metrics can be seen in Figure 1.

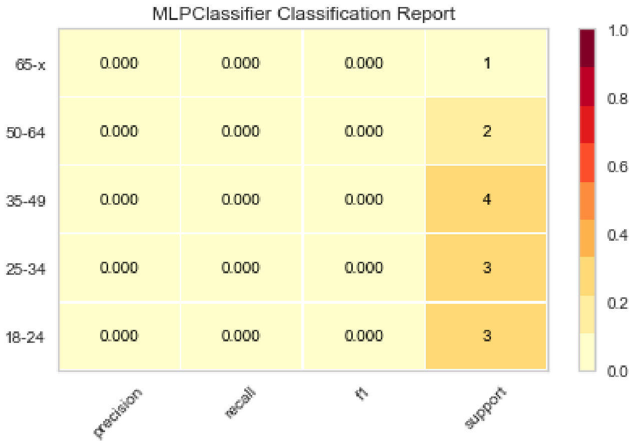

Fig. 1. Metrics for Baseline Age Classifier

The accuracy of the trained age classifier showed improvement, with an accuracy of approximately 54%. Figure 2 demonstrates the results for the other metrics.

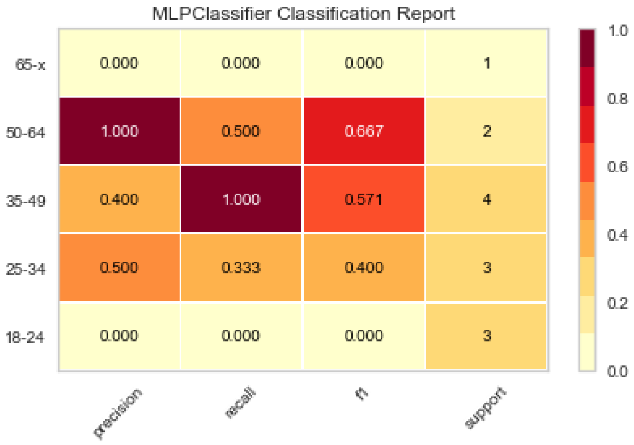

Fig. 2. Metrics for Trained Age Classifier

The dummy classifier's baseline accuracy was around 16% for time period identification. The other metrics are shown in Figure 3.

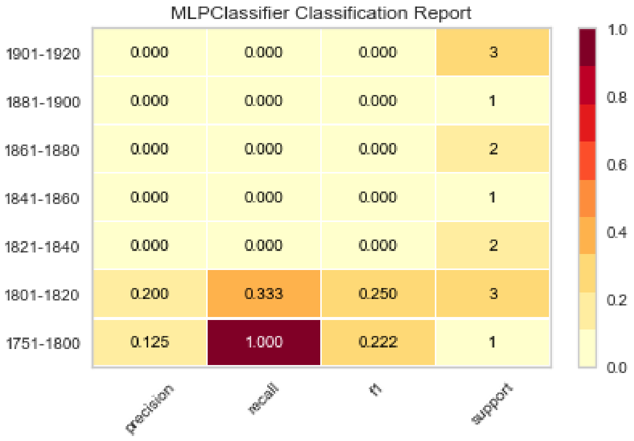

Fig. 3. Metrics for Baseline Time Period Classifier

Like the age classifier, the trained time period classifier also improved, with an accuracy of 46% on the test data. The precision, recall, and F1 are shown in Figure 4.

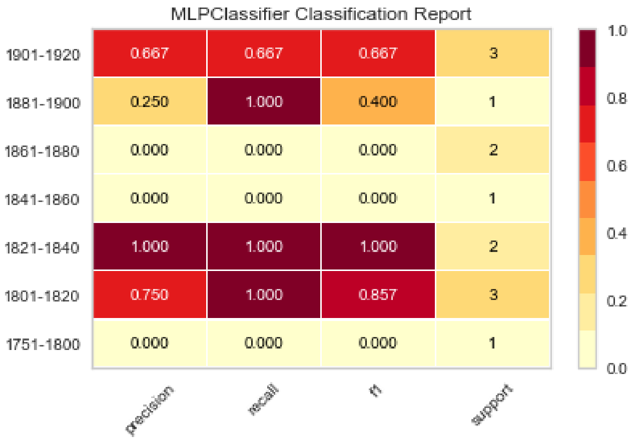

Fig. 4. Metrics for Trained Time Period Classifier

Based on the overall results above, it is clear that the features chosen for the classifiers had a positive impact in predicting the correct age and time period ranges. However, the features with the most impact are unidentifiable from this information; the Random Forest Classifier in the next section sheds more light on the features with the highest significance.

5.1.2 Random Forest. The dataset which we used to train the Random Forest Classifiers was significantly larger than that for the Multilayer Perceptrons. The dataset was comprised of 380 books written by a variety of authors. 80% of the dataset (approx. 304 books) was used for training, while 20% (approx. 76 books) was used for testing. With the use of the Random Forest model, an additional metric was introduced to evaluate feature importance. Figure FILL and Figure FILL below display the top ten most important features for the age classifier and the time period classifier respectively. The blue bars for each feature in the top ten represents its percentage importance relative to the most important feature (at the top of the graph).

The baseline accuracy for the dummy age classifier was 21%. The other metrics are low as well, shown in Figure 5.

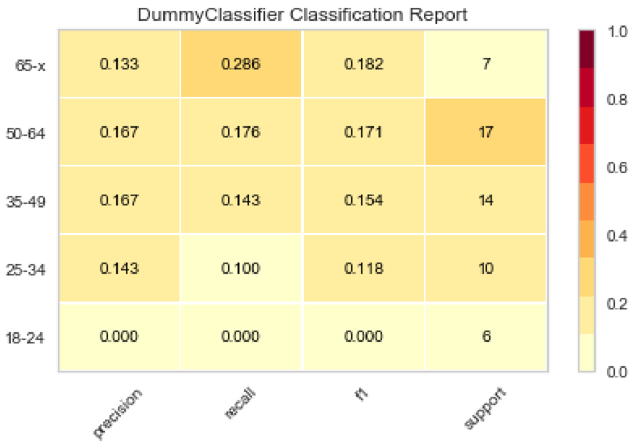

Fig. 5. Metrics for Baseline Age Classifier

Perhaps as a result of the larger dataset and modified feature set, the accuracy of the trained age classifier, while greater than the baseline (at 34%), was less than that of the Multilayer Perceptron accuracy. The other metrics, shown below in Figure 6, were also only a slight improvement from the baseline.

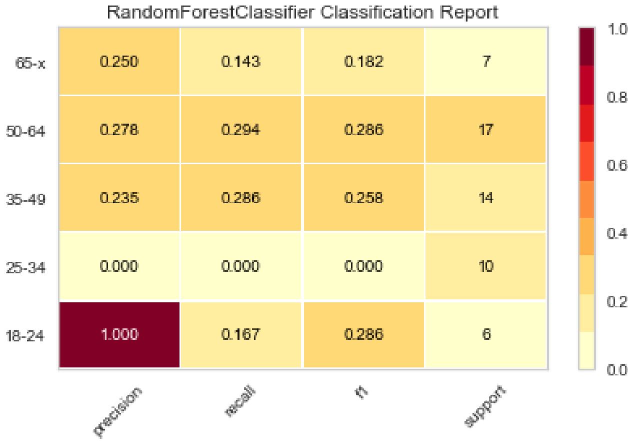

Fig. 6. Metrics for Trained Age Classifier

The most importance features in classifying age are illustrated in Figure 7. It appears that unigram features had the most impact, with the word *whole* having the highest significance and *though* coming in second place. The labels *w2v* indicate word2vec features; with the nature of our word2vec implementation, there is no way to determine what specific word senses are represented by *w2v*. Another aspect of the figure to note is that after the top eight values, those in ninth and tenth place apparently have little to no impact.

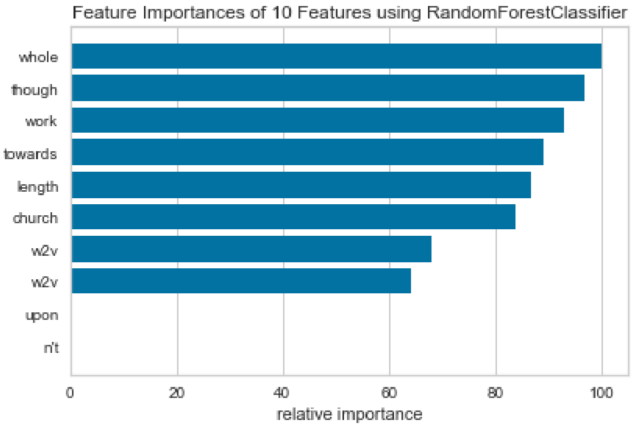

Fig. 7. Top 10 Features for Trained Age Classifier

The results of time period classification were consistent with those of age classification in that the accuracy of the trained model (32%) was greater than that of the dummy classifier (23%) but less than the accuracy of the neural network on a smaller dataset. Figure 8 displays the metrics of the dummy classifier and Figure 9 shows those of the trained classifier.

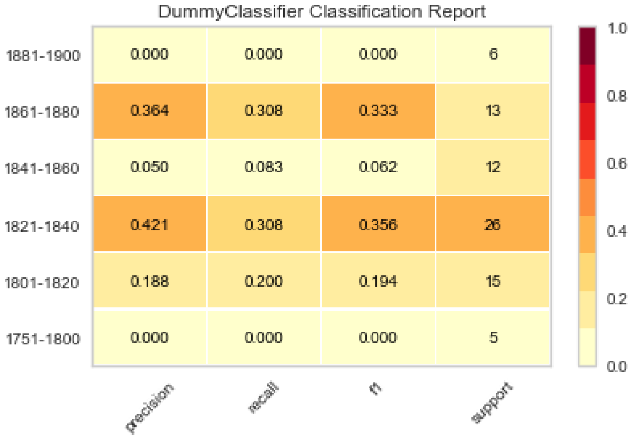

Fig. 8. Metrics for Baseline Time Period Classifier

Figure 10 displays the most significant features for the trained time period classifier. Like the age classifier, the top two features are unigrams: *where* and *does*. Unlike the features for the age classifier, a part-of-speech tag *WP* was considered important as well.

5.2 Training Speed and Performance

Regardless of the model, feature extraction took the most amount of time. Part of the difficulty in having a larger dataset was how slow the process was to form feature vectors to be passed to the models.

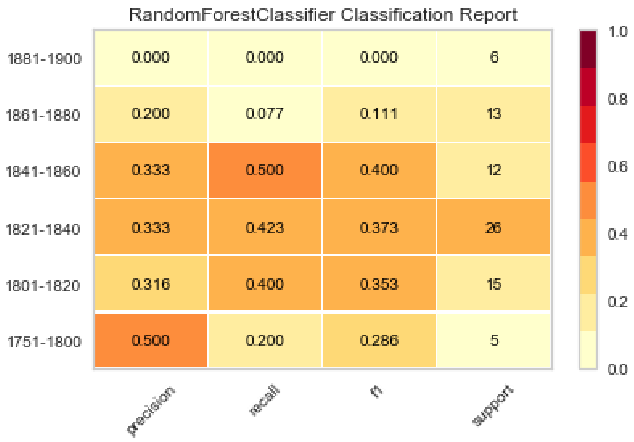

Fig. 9. Metrics for Trained Time Period Classifier

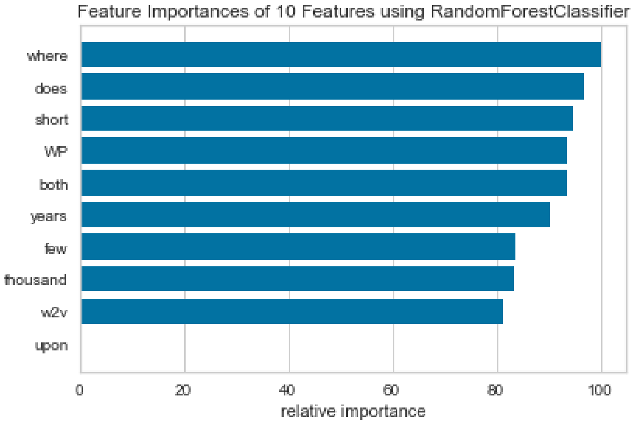

Fig. 10. Top 10 Features for Trained Time Period Classifier

In terms of model training, the cross-validation and training of the Multilayer Perceptron was slower than that of the Random Forest classifier, even with different dataset sizes.

6 CONCLUSION

For this project, we experimented with different feature and model combinations to create age and time period classifiers for novels written in the 18th and 19th centuries. In the future, we would like to experiment more with feature dropping

REFERENCES

[1] Marco Dinarelli1 Anne Garcia-Fernandez, Anne-Laure Ligozat and Delphine Bernhard. 2011. When Was It Written? Automatically Determining Publication Dates. *String Processing and Information Retrieval Lecture Notes in Computer Science* (2011). https://perso.limsi.fr/annlor/docs/spire11.pdf

[2] Noah A. Smith Dong Nguyen and Carolyn P. Rosé. 2011. Author Age Prediction from Text using Linear Regression. *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2011), 115–123. hhttp://delivery.acm.org/10.1145/2110000/2107651/p115-nguyen.pdf?ip=71.174.251.115&id=2107651&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1552696258_041e7e8ac84ce4b4aaf78893d9f88543

[3] Michael Gamon. 2004. Linguistic correlates of style. *Proceedings of the 20th international conference on Computational Linguistics - COLING 04* (2004). https://doi.org/10.3115/1220355.1220443

[4] Foaad Khosmood and Robert Levinson. 2010. Automatic Synonym and Phrase Replacement Show Promise for Style Transformation. *2010 Ninth International Conference on Machine Learning and Applications* (2010). https://doi.org/10.1109/icmla.2010.153

[5] Paolo Rosso. 2017. Author Profiling at PAN: from Age and Gender Identification to Language Variety Identification (invited talk). *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (2017). https://doi.org/10.18653/v1/w17-1205