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 (February 2019), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As natural language processing technologies have improved, there has been an increased interest in utilizing them to analyze and categorize literary works. From determining gender to author's native language to genre, NLP has been used to not only classify literature, but to determine which features are most salient in said classification.

We hope to add to this growing body of work by analyzing two elements that have sparse literature — author's age and time period of writing. This information can be applied to a wide variety of problems in literary analysis, such as dating a piece of unknown origin or examining how age affects an author's writing style over their lifetime. Our goal is to create a model that, given a work of literature written between 1700 – 1920, will output the author's age, the decade in which the work was written, and the features that determined these results.

2 RELATED WORK

This section discusses previous work associated with author age identification and time period identification. Before we begin, we must note that our research is unique in that it addresses both problems 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 might lead to alternative feature vectors or a different form of analysis.

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

Nguyen et. al. (2011) [4] 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) [7] 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) [3] 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) [6] 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) [5] also discussed style classification like Khosmood and Levinson (2008) [6], 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.

3 DATASETS

The corpus that we will use to retrieve data is Project Gutenberg [1], a digital library organized by volunteers that contains more than 58,000 books in the public domain. From Project Gutenberg's collection we will choose several works that fulfill our necessary criteria. The requirements for the works are defined as follows: they must be novels, they must be originally published in English, and

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/2019/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

both the author and the publication year of each work must be known. Every publication year will be within the range of 1700-1920, as novels published within this time period are in the public domain and are more likely to be found in the digital library. The author of each novel must have been alive at the time of publication, his/her age at the time of publication must be known and each author should be "notable" in the sense that his/her novels were widely read during his/her lifetime or posthumously. Examples of authors who would fulfill the aforementioned standards include Jane Austen, Charles Dickens, and Mary Shelley.

Since novels for training and testing have yet to be chosen, a specific size for the data is not yet defined. However, as Project Gutenberg has over 58,000 books to offer, we have much flexibility in determining the ultimate size of the dataset we will use. The information contained within the Project Gutenberg data (for an individual novel) includes the name of the author, the year of publication and the entire work of the author in plain-text (and in additional formats). Supplementary details about an author's age when his/her books were published will likely be retrieved from Wikipedia [2].

We determined that Project Gutenberg would be the most appropriate corpus for our endeavor for a number of reasons. The digital library contains one of the most extensive selections of e-books under the public domain that we could find online. We can analyze any book from this collection without concern about copyright infringement. The necessary publication and author information is easily accessible as well. Finally, every work is available as a plain-text document, which will allow us to easily strip away inessential metadata and process the actual content in its completion.

4 METHODOLOGY

Both age determination and time determination are tasks that require looking at multiple features. Typical classification features, such as POS and N-grams, should be considered. We can also consider features that are useful for these particular problems, such as named entities and neologisms. Previous work on identifying time period has utilized the "date of birth" categories on Wikipedia, as well as Google Books data on neologisms and anarchism to identify likely year of publication (Garcia-Fernandez et. al. 2011 [3]). We will continue to examine related literature for features to experiment with and determine which features are most critical to our task in the training stage.

We will be training two separate models, one on author age and one on time frame, and then combining the outputs. Our training process will be supervised, as the output only requires a time-frame and an age range for the author. These are attributes that can be discovered and annotated relatively easily, although it will take effort to look up author age, as such information is not readily included in corpus data. As such, our dataset for the author age model could be notably smaller than that for the time period model.

As we have a significant number of features to consider, we will be using a probabilistic model. We can determine which model to use based on related work. Nguyen et. al.'s (2011) [4] work on author age determination utilized a regression-based model, whereas Garcia-Fernandez et. al. (2011) [3] and Rangel and Rosso (2013) [7] utilized an SVM approach. More research will be needed to determine which of these is better for our task, or if it would be better to use a different model entirely.

5 EVALUATION

We will evaluate our approach using a test set (20% of the total number of books, and 80% for the development set). While extrinsic evaluation would likely be more accurate, there is currently no extrinsic task that we could use to evaluate our model, as our model is extremely specific. Thus, we will be using intrinsic evaluation, where the model is applied to the test set of novels in which the age of author and year of publication are not provided. We will compute the "goodness" of our models' prediction quantitatively, using the following formula:

$$goodness = \frac{\tilde{predicted} - \tilde{actual}}{range} \quad (1)$$

Based on the formula, the "goodness" will be normalized to a number between 0 and 1. The smaller the number, the better the model.

We will also compute the precision, recall, and accuracy for each age range, time period, and each author, where a true positive is defined as predicted age/time period matching the actual age/time period. The F-measure or F-score can also be computed to give a combined measure that assesses the precision and recall trade-off.

REFERENCES

[1] [n. d.]. Project Gutenberg. Retrieved February 07, 2019 from http://www.gutenberg.org/

[2] 2019. List of literary movements. Retrieved February 07, 2019 from https://en.wikipedia.org/wiki/List_of_literary_movements

[3] 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

[4] Noah A. Smith Dong Nguyen and Carolyn P. Rosé. 2011. Author Profiling at PAN: from Age and Gender Identification to Language Variety Identification (invited talk). *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (2011), 115–123. http://delivery.acm.org/10.1145/2110000/2107651/p115-nguyen.pdf?ip=155.33.135.5&id=2107651&acc=OPEN&key=AA86BE8B6928DDC7%2EC2B8A117C7A71F5A%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1549501746_e14634bd32267eed456ac68bc06e16fd

[5] 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

[6] 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

[7] 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