Author Identification with POS Tagging

Katherine Schulz
Georgetown University
Analytics Department
Master's Candidate
ks1533@georgetown.edu

Abstract

This project tested the efficacy of POS tagging for author identification. From a dataset of 1,605 texts written by 134 authors collected from The Project Gutenberg repository, I built an SVM and a bigram CNN to perform author identification on tokenized texts and POS tags. Over two experiments, one in individual author classification, and one in binned author classification, I found that models trained on POS tags can successfully identify authors, but not as well as models trained on tokens.

1 Introduction

Author identification has applications in digital forensics, anti-terrorism, and anti-plagiarism where it assists investigators in tracking the work of specific people (Stamatatos, 2009). Author identification relies primarily on the "extraction of stylometric features" that represent an author's general writing style, and models tends to perform better when trained on larger samples of an author's text (Markov, Stamatatos, & Sidorov, 2017; Qian, He, & Zhang, 2017).

This project deals with author identification using a cross-topic subset of the Project Gutenberg repository. Specifically, I trained a support vector machine (SVM) and a bigram convolutional neural network (CNN) model to perform author identification on tokenized texts and part of speech (POS) tags. I performed two experiments: the first experiment compares the classification accuracy rates of the baseline SVM model to the bigram CNN, each training on various input lengths of tokens and POS tags; the second experiment examines the classification accuracy rates of the bigram CNN trained on tokens and POS tags when the authors are binned based on Ward's method of clustering.

2 Related Work

Author identification relies on extracting and training models on one or more of the three main types of stylometric features: lexical, syntactic, and content-specific. Lexical features may include tokenizing text and examining character-based or word-based patterns, while syntactic features may include POS tagging, sentence structure, or word counts at the sentence level. Content-specific features may include words, phrases, or writing structures based on domain expertise in a specific topic (Ghanem, El-Makky, & Mohsen, 2016).

Prior to the wide-spread use of deep learning, researchers typically performed author identification with SVM classifiers, which were successful when trained over longer documents, but unsuccessful on shorter ones (Qian, He, & Zhang, 2017). Deep learning models have since replaced SVM classifiers, as deep learning models can uncover multiple layers of stylometric features without supervision (Ghanem, El-Makky, & Mohsen, 2016).

Successful approaches to author identification using the Project Gutenberg repository include a publication on Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) networks trained on word embeddings initialized with Global Vectors for Word Representation (GloVe). Qian et al. (2017) found that article-level models

attained higher accuracy rates (between 60 and 70 percent) than the sentence-level models (between 40 and 45 percent) (Qian, He, & Zhang, 2017).

Author identification models may be enhanced by discourse embeddings, which can track the "grammatical relations of salient entities" such as subjects, objects, and verbs, across sentences. Ferracane et al. (2017) demonstrated that local and global discourse embeddings improved CNN accuracy rates over a baseline SVM by 2 to 6 percent, although global discourse embeddings generally performed better than local discourse embeddings. Additionally, Ferracane et al. (2017) found that "adding any discourse information improves [author identification] consistently on longer documents but has mixed results on shorter documents" (Ferracane, Wang, & Mooney, 2017).

To deal with cross-topic texts, another publication successfully demonstrated that pre-processing techniques, such as replacing digits and named entities with placeholder symbols, improved classification accuracy rates by about 10 percent on several SVM and multinomial naïve Bayes (MNB) classifiers. The publication notes, however, that in the case of single-topic modeling, the pre-processing does significantly improve the models' accuracy rates (Markov, Stamatatos, & Sidorov, 2017).

3 Dataset

The dataset for this project came from a collection of texts pulled from the Project Gutenberg repository that contains 3,036 works by 142 authors in the English language. The collection includes fiction and non-fiction books, essays, short stories, speeches, and poems written from the 17th to 20th centuries. This collection has been used previously in a publication to study the properties of word collocation networks across different genres (Lahiri, 2014). It has already been "manually cleaned to remove metadata, license information, and transcribers' notes, as much as possible" (Lahiri, 2014).

4 Methodology

I processed the data and designed the models to conduct two experiments on the efficacy of POS tags in author identification. The computational expense of the data processing forced me to subset my initial dataset. The models were then trained and tested using texts from the subset.

4.1 Data Processing

For this project, I subsetted the initial data for computational efficiency. Issues with computational efficiency stemmed from the Stanford POS Tagger. On a graphics processing unit (GPU) runtime host through Google Colaboratory, I encountered Java heap space memory errors, even when I maximized the available heap space to 4,096 megabytes (MB). These errors forced me to reduce the initial dataset by about half. From the initial dataset of 3,036 works by 142 authors, I took 1,605 works by 134 authors. Table 1 below summarizes the final dataset.

Table 1: Description of Final Dataset

Dataset Summary	
Number of Texts	1,605
Number of Authors	142
Average words per sentence	22
Average tokens per author	70,508

To process the texts, I replaced all digits in the texts with zeros to mitigate the effects of topic-specific numbers based on the findings of Markov et al. (Markov, Stamatatos, & Sidorov, 2017). I then used the NLTK tokenizer and Stanford POS Tagger *left3words* model to process the text. I chose the *left3words* model over the *bidirectional* model because *left3words'* runtime is an order of magnitude faster (The Stanford Natural Language Processing Group, n.d.).

To bin the authors, I used the *sklearn TfidfVectorizer*() function to vectorize the texts and performed Ward's method using cosine distances. I set the number of bins to four to capture roughly equal numbers of authors within each bin. The full dendrogram produced by Ward's method of clustering is located in Appendix A.

4.2 Models

The models were trained and tested on four sets of tokens and POS tags of lengths: 100, 500, 1,000, and 5,000. I chose these lengths to represent approximately one paragraph, one page, two pages,

and ten pages of an author's work. I split the data randomly into an 80 percent training set and a 20 percent test set. I fit two models on the data: an SVM (to serve as a baseline) and a bigram CNN. To build the SVM, I vectorized the tokens and POS tags with the *sklearn TfidfVectorizer*(). For the bigram CNN, I set the maximum number of features at 25,000 for efficiency purposes. The bigram CNN trained for 50 epochs with an *RMSprop* optimizer.

5 Results

I conducted two experiments with the models: the first experiment tested the classification accuracy rates for individual authors based on tokens and POS tags; the second experiment tested the classification accuracy rates for binned authors based on tokens and POS tags. Overall, both models performed better than random chance in individual author and binned author classification on both tokens and POS tags.

5.1 Experiment 1: Individual Authors

In the first experiment, the SVM and bigram CNN models were tested on classification accuracy for individual authors given different input lengths of POS tags and tokens. Table 2 below shows the accuracy rates for each model on the POS tags and tokens of varying input lengths. The SVM and bigram CNN had comparable accuracy rates. Token accuracy ranged between 0.673 and 0.860 for both models, while POS tag accuracy ranged between 0.125 and 0.170. Increasing input length did not appear to improve accuracy rates overall. For tokens, both the SVM and bigram CNN performed best on an input length of 500. For POS tags, the SVM performed best on an input length of 5,000, while the bigram CNN performed best on an input length of 100.

Table 2: Individual Author Classification Accuracy

Input Length	POS Tag Accuracy		Token Accuracy	
	SVM	Bigram CNN	SVM	Bigram CNN
100	0.125	**0.171**	0.754	0.673
500	0.156	0.153	**0.86**	**0.798**
1,000	0.125	0.134	0.798	0.782
5,000	**0.168**	0.146	0.704	0.717

Figure 1 below show the validation accuracies for the bigram CNN for varying input lengths of tokens and POS tags. The figure indicates that over 50 epochs, improvements in the accuracy rates do not correlate with increasing input lengths. For the tokens, validation accuracies for all input lengths appear to plateau at about 25 epochs, whereas for the POS tags, validation accuracy appears to continue improving up to 50 epochs. Validation accuracy on this dataset with POS tags may improve if the bigram CNN is trained over more than 50 epochs.

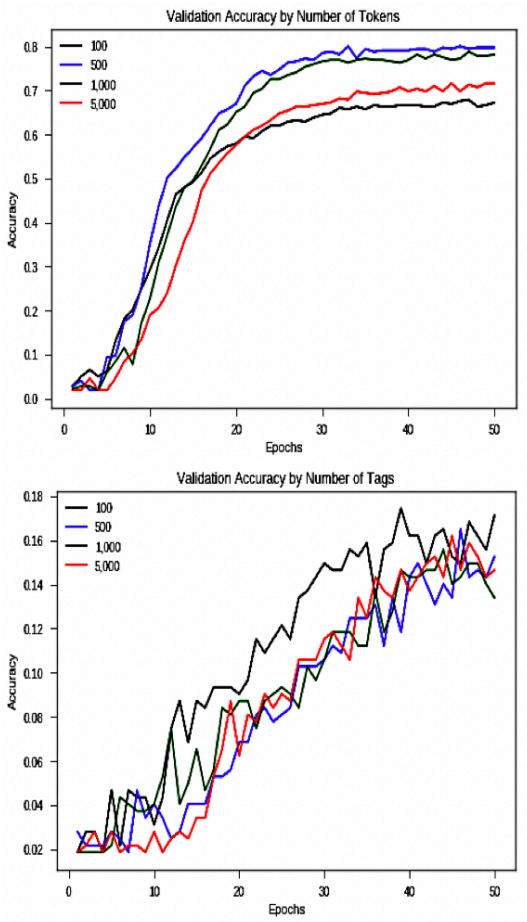


Figure 1: Validation Accuracies for Bigram CNN

5.2 Experiment 2: Binned Authors

In the second experiment, the bigram CNN model was tested on classification accuracy for binned authors given the input lengths that performed best in Experiment 1. Figure 2 below shows the validation accuracies for the bigram CNN on 500

tokens and 100 POS tags. Binning classification accuracies for both the tokens and POS tags appears to level off by about 10 epochs. Overall, the bigram CNN reached a bin classification accuracy of about 0.928 on 500 tokens and 0.657 on 100 POS tags. The improvements in binning classification accuracies over individual author classification accuracies could reduce the original set of 134 potential authors to a subset of approximately 34 authors with accuracy rates of 66 to 93 percent.

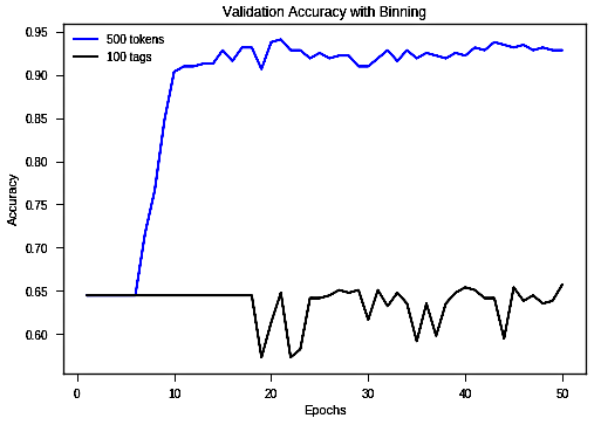


Figure 2: Validation Accuracies for Bigram CNN with Binning

6 Conclusion

This project examined the efficacy of POS tagging for author identification. The data comprised of a subset of the Project Gutenberg repository, including 1,605 works from 134 authors. I trained an SVM and bigram CNN on various lengths of tokens and POS tags. I found that the deep learning model (bigram CNN) did not improve accuracy rates over the supervised machine learning model (SVM). The models trained on input lengths of 100, 500, 1,000 and 5,000 tokens and POS tags; however, increasing the input lengths did not improve classification accuracies overall. The first experiment showed that POS tags could identify authors with an accuracy rate of up to 17 percent. The second experiment showed that POS tags could reduce a pool of potential authors by three-quarters with an accuracy rate of up to 66 percent. The results of these models may not generalize well, as the data comprised of high-quality writing samples from professional authors and scholars.

References

Ferracane, E., Wang, S., & Mooney, R. J. (2017, December 1). Leveraging Discourse Information Effectively for Authorship Attribution. Retrieved November 2018, from Proceedings of the The 8th International Joint Conference on Natural Language Processing: http://aclweb.org/anthology/I17-1059

Ghanem, N., El-Makky, N. M., & Mohsen, A. M. (2016). Author Identification using Deep Learning. Retrieved November 2018, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7838265

Lahiri, S. (2014, April). Complexity of Word Collocation Networks: A Preliminary Structural Analysis. Retrieved November 2018, from Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics: http://www.aclweb.org/anthology/E14-3011

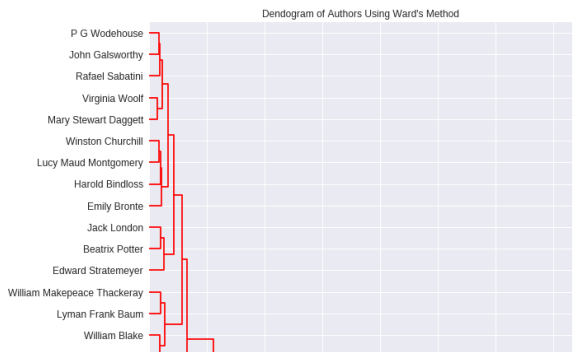
Markov, I., Stamatatos, E., & Sidorov, G. (2017, April). Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. Retrieved November 2018, from http://www.cic.ipn.mx/~sidorov/CICLing-Markov-Preprint.pdf

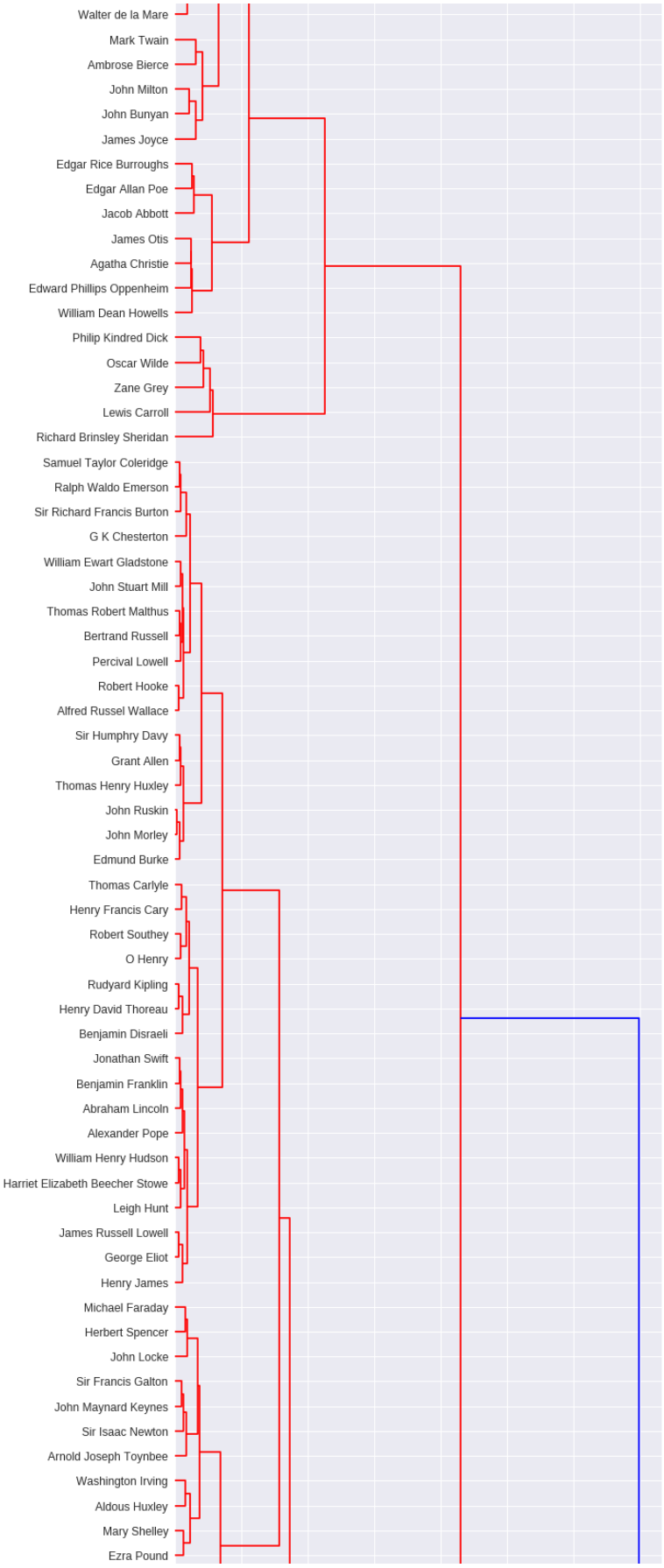
Qian, C., He, T., & Zhang, R. (2017). Deep Learning based Authorship Identification. Retrieved November 2018, from https://web.stanford.edu/class/cs224n/reports/2760185.pdf

Stamatatos, E. (2009, March). Retrieved November 2018, from A Survey of Modern Authorship Attribution Methods: https://pdfs.semanticscholar.org/d25c/27c7a3e9f41f150e8eadbad34c1c05d67510.pdf?_ga=2.101008001.1800055236.1543592516-1748937950.1543592516

The Stanford Natural Language Processing Group. (n.d.). Stanford POS tagger FAQ. Retrieved December 2018, from https://nlp.stanford.edu/software/pos-tagger-faq.html

Appendix A. Dendrogram of Ward's Method of Clustering




Walter de la Mare
Mark Twain
Ambrose Bierce
John Milton
John Bunyan
James Joyce
Edgar Rice Burroughs
Edgar Allan Poe
Jacob Abbott
James Otis
Agatha Christie
Edward Phillips Oppenheim
William Dean Howells
Philip Kindred Dick
Oscar Wilde
Zane Grey
Lewis Carroll
Richard Brinsley Sheridan
Samuel Taylor Coleridge
Ralph Waldo Emerson
Sir Richard Francis Burton
G K Chesterton
William Ewart Gladstone
John Stuart Mill
Thomas Robert Malthus
Bertrand Russell
Percival Lowell
Robert Hooke
Alfred Russel Wallace
Sir Humphry Davy
Grant Allen
Thomas Henry Huxley
John Ruskin
John Morley
Edmund Burke
Thomas Carlyle
Henry Francis Cary
Robert Southey
O Henry
Rudyard Kipling
Henry David Thoreau
Benjamin Disraeli
Jonathan Swift
Benjamin Franklin
Abraham Lincoln
Alexander Pope
William Henry Hudson
Harriet Elizabeth Beecher Stowe
Leigh Hunt
James Russell Lowell
George Eliot
Henry James
Michael Faraday
Herbert Spencer
John Locke
Sir Francis Galton
John Maynard Keynes
Sir Isaac Newton
Arnold Joseph Toynbee
Washington Irving
Aldous Huxley
Mary Shelley
Ezra Pound


T S Eliot
P B Shelley
Stephen Leacock
Charles Darwin
Albert Einstein
William Wordsworth
William Penn
Charles Kingsley
Lord Tennyson
John Keats
Elizabeth Barrett Browning
William J Long
Louisa May Alcott
James Matthew Barrie
Robert Louis Stevenson
Howard Pyle
Sinclair Lewis
Sir Walter Scott
James Fenimore Cooper
George Alfred Henty
Herman Melville
Robert Burns
Charles Dickens
Ulysses Grant
Herbert George Wells
Charlotte Bronte
William Butler Yeats
Lord Byron
R M Ballantyne
Joseph Conrad
Bram Stoker
Bret Harte
Robert Frost
Daniel Defoe
Hamlin Garland
D H Lawrence
John Dryden
Henry Rider Haggard
Walt Whitman
George Bernard Shaw
Wilkie Collins
Andrew Lang
Baronness Orczy
Thomas Hardy
Sir Arthur Conan Doyle
Nathaniel Hawthorne
Hector Hugh Munro
Frank Richard Stockton
Charlotte Mary Yonge
Anthony Trollope
Jane Austen
William Wymark Jacobs
Jerome Klapka Jerome
Robert Browning
Eugene O
William Somerset Maugham
Thornton Waldo Burgess
0
2
4
6
8
10
12
14