Spooky Author Classification: Re-thinking Authorship Attribution

Patrick Prioletti
Syracuse University
pjpriole@syr.edu

1 Abstract

While authorship classification is typically oriented around style and structural aspects of the author, this paper will review the use and comparison of a few combinations of content features, structural features, semantic features, and stylistic features in two common machine learning models. This paper will bring some skepticism to the current state of contemporary authorship classification literature as we find that including text content features appears to outperform any other combination of features tested.

2 Introduction

This analysis is a classification task of text documents from a variety of three famous horror/suspense novelists: Edgar Allen Poe, Mary Shelley, and HP Lovecraft. Correctly identifying the authorship of a sentence is the primary task to be completed, however, to better compare the model behavior we will look into the specifics of the classification errors and decisions as well as feature importance coefficients. Seen in Table 1 are the class/author observation distributions, where we can see the classes are imbalanced. For a performance baseline measure, we expect a "good classifier" to achieve an accuracy score of at least 40.35% - the equivalent of voting for the majority class on every observation. Authorship classification has been reportedly achieved by using style and structure measurements as opposed to the content, which has typically been viewed as noise, while aspects of language use such as function words and other style characteristics are better for performing authorship classification [1] [2] [3]. These data, as opposed to a complete page of the books produced by their respective authors, consists of sentences of any given page by the authors' works.

Author	Count
Edgar Allen Poe	7900
Mary Shelley	6044
HP Lovecraft	5635
Total	19579

Table 1

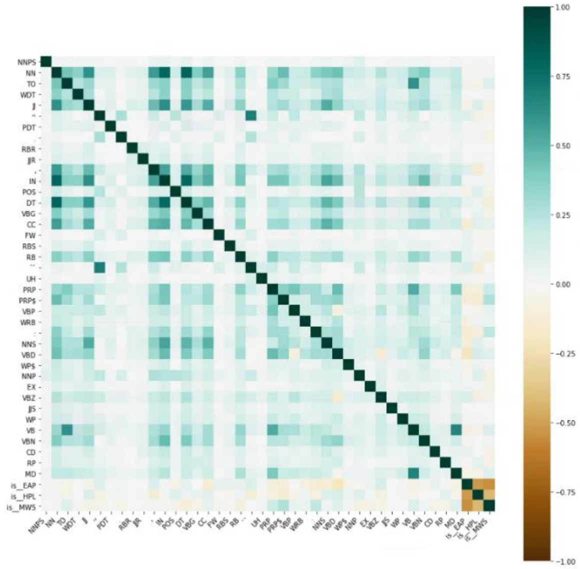
2 Methods

Exploring the space feature performance of authorship classification on this corpus, we utilize multiple machine learning algorithms and natural language processing (NLP) techniques for purposes of feature production and for the classification of the documents. To encapsulate and test various measurements of authorship attribution, we use part-of-speech (POS) tagging to extract the count of words per document of grammar categories to measure structure and style. Latent Dirichlet Allocation (LDA) is utilized to get each documents' distribution across some topics generated by the algorithm. Function words are counted, and the counts are stored as features in one instance. Finally, the text is vectorized into a term-frequency inverse-document-frequency (TF-IDF) matrix.

Subsequently, we utilize a Multinomial Naïve Bayes (MNB) classifier as well as a Support Vector Machine (SVM) classifier to perform the classification task of authorship attribution to the documents. As will be discussed later, the modeling process utilizes different combinations of features to observe the differences in performance, however, each instance of evaluation utilized the same training and testing observations to complete the classification, therefore making performance comparison very informative.

2.1 POS Tagging

For identifying and counting POS tags, we utilize NLTK's pos_tag function, which utilizes the averaged_perceptron_tagger function. Before the

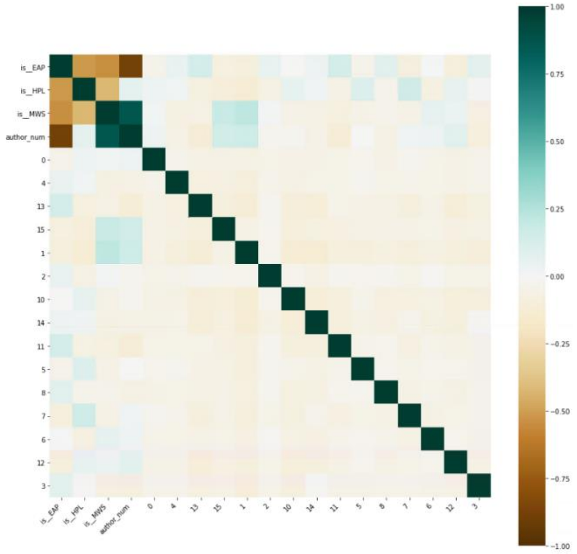


tagger produces the document tag set, we tokenize the documents with the word_tokenize function from NLTK. This allows the tagging process to produce the POS tag for each given word, and allows the tags to be counted and stored for each document as a feature. Figure 1 displays the correlation of each feature, including three Boolean values of whether the document belongs to the author or not. We can see that there seems to be a few interesting correlation values between the tag feature and a few of the authors, most notably, Mary Shelley's documents seem to be more positively correlated with POS features that HP Lovecraft and Edgar Allen Poe are not. This is promising as for the expected features which are typically heeded to be good predictors of authorship.

2.2 LDA Modeling

LDA is an unsupervised generative algorithm introduced by Blei et. Al. in 2003 [4], often referred to as topic modeling, as it is proficient to extract probabilistic distributions and may be used for mapping semantic structures of corpora. This analysis utilizes LDA to extract the topic distributions of documents given some generated topic distribution. In this context, LDA was implemented using the Gensim project package wrapper for Python. The specific implementation is based on a variant of LDA [5] and utilizes an evaluation metric called C_v coherence introduced by Röder et. Al. [6] to enforce an optimal K value for the algorithm to produce. To validate and inspect this topic distribution, we rely on a principal component projection of the topic marginal distributions across the corpus introduced by Sievert and Shirley [7] with the PyLDAvis package.

After producing a model with a K value of 16 with a C_v coherence value of 0.372, the marginal topic distributions of each document's topics are stored as features to be utilized for the classification task. Figure 3 is the correlation values of the document topic distributions in a similar framing to that of the POS tags. We can observe that there are some promising correlations with the three Boolean features representing the author and the document topic features produced.



2.3 TF-IDF Text Vectorization

Scikit-learn's Python module was used for text vectorization to extract content characteristic features. A few different vectorizers were tested, however, using TF-IDF appears to be the best option, especially as the value produced smooths the metrics biasing towards unique vocabulary. The vectorizer produces both unigram and bigram counts to capture some aspects of co-location and to avoid overfitting we require the inclusion of vocabulary to occur at least three times across all documents, which yields 13,883 features across the training set of 15,663. Alternatively, a minimum document frequency of two yields ~25,600 features while one or no minimum document frequency criteria crashed the development environment. Text features such as this, which focus on content, are not typically shown to be as effective as style or structure features for authorship attribution, nonetheless we test its efficacy in this context to strive towards a better classification performance.

2.4 Function Words

Function words gathered from Jim O'Shea's personal website, based off of his use of these function words in classification tasks [8]. Function words are the use of various words which are typically seen as noise, however, are seemingly useful when dealing with an authorship attribution task [1] [2] [3]. While the function words list here is not likely to be all-inclusive, this is the largest list of function words that could be found and made readily available. Additionally, this function word list is a list of English words which typically indicate a function within a sentence and will not pick up other languages and may add some confusion to the analysis if a word in the list can also be used as an object or subject word.

3 Modeling

As discussed previously, modeling the data serves two purposes: the first being to classify the documents and the second being to observe inferences and gain insights into feature-to-class relationships. Two algorithms were deployed, evaluated, and compared to gain a better understanding of the language patterns, writing style patterns and other linguistic structural attributes which are indicative of each of the three authors. First, a MNB classifier is used and tested with different subsets of features, then a linear SVM is used in similar fashion with the same subsets of features. MNB utilizes Bayesian probability estimates, where we estimate the probability of class given some attributes and predict this class with the following notation:

$$P(x_i|y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Where $N_{yi} = \Sigma_{x \epsilon T^{x_i}}$ being the number of times feature i appears in a sample of the class y in the training set T and $N_y = \Sigma_{i=1}^{n} N_{yi}$ is the total count of all features of class y. For the purposes of this analysis, we do not pay attention to the smoothing parameter α.

Similarly, SVM - more specifically, scikit-learn's linearSVC - makes a decision by drawing a hyperplane, or multidimensional linear surface to separate the classes linearly. While the formulation of these decision boundaries is conceptually understandable, we will ignore the notation, as the details of the derivation of a multi class classification function is beyond the scope of the task since we do not optimize our C value to simplify our interpretation of the results.

3.1 Multinomial Naïve Bayes

Two sets of features yield different performance measures and seem to be counter-intuitive of the current authorship classification literature. In Table 2 and Table 3 we can see the two different performance metrics of the two feature sets used to train the model. Optimal performance was achieved using text data alone as opposed to using structure and style indicators to increase authorship classification. Moreover, we can see in Table 4 and Table 5, the performance of using only style and structure characteristic features alongside function words yields the worst performance of all combinations, which is evidence against the literature consensus of best authorship attribution techniques. We can see the log probability of the top ten features for each author for the given algorithms in Table 6 and Table 7, where we can see that the POS and LDA attributes are disproportionately highly indicative of authorship for each class, however, the model performance is worse when using these features.

These findings are understandably confusing, as authorship typically considers the introduction of content characteristics as noise, but we can see that in our case, the opposite is true, meaning the introduction of style, structure, and function word characteristics are behaving as the noise being introduced. To ensure that this is not anomalous behavior we observe a second model, SVM, and carry out an analysis mirroring the process used to classify authorship utilizing MNB with different feature sets.

Table 2

	MNB TF-IDF, POS, LDA			MNB TF-IDF			
	Precision	Recall	F1	Precision	Recall	F1	Support
0	0.77	0.79	0.78	0.78	0.87	0.82	1570
1	0.77	0.69	0.73	0.87	0.74	0.80	1071
2	0.73	0.76	0.75	0.83	0.82	0.82	1275
Accuracy			0.75			0.82	3916
Macro Average	0.75	0.75	0.75	0.83	0.81	0.81	3916
Weighted Average	0.76	0.75	0.75	0.82	0.82	0.82	3916

Table 3

	MNB TF-IDF, POS, LDA			MNB TF-IDF			
Actual	EAP	MWS	HPL	EAP	MWS	HPL	Totals
EAP	1243	124	203	1360	78	132	1570
MWS	180	740	151	198	740	79	1071
HPL	199	103	973	187	45	1043	1275
Totals	1622	967	1327	1724	863	1254	3916

Table 4

	MNB POS, LDA, Func. Words			
	Precision	Recall	F1	Support
0	0.69	0.64	0.66	1570
1	0.57	0.63	0.60	1071
2	0.60	0.61	0.60	1275
Accuracy			0.63	3916
Macro Average	0.62	0.63	0.62	3916
Weighted Average	0.63	0.63	0.63	3916

Table 5

	MNB POS, LDA, Func. Words			
Actual	EAP	MWS	HPL	Totals
EAP	1000	264	306	1570
MWS	188	678	205	1071
HPL	261	241	773	1275
Totals	1499	1183	1284	3916

Table 6

MNB TF-IDF, POS, LDA					
Edgar Allen Poe		Mary Shelley		HP Lovecraft	
-6.075	RP	-6.081	Topic 14	-6.215	Topic 12
-6.011	Topic 14	-5.974	Topic 10	-5.978	Topic 15
-5.860	WRB	-5.924	“‘’”	-5.867	CD
-5.837	Topic 13	-5.787	WP	-5.811	“‘’”
-5.822	WP	-5.748	RP	-5.772	WP
-5.354	CD	-5.659	POS	-5.602	Topic 1
-5.234	‘:’	-5.612	CD	-5.588	WRB
-5.150	‘``’	-5.496	WRB	-5.473	WDT
-5.133	“‘’”	-5.286	VBZ	-5.101	VBZ
-5.089	WDT	-5.208	‘:’	-5.096	VBG

Table 7

MNB TF-IDF					
Edgar Allen Poe		Mary Shelley		HP Lovecraft	
-7.064	Within	-7.119	See	-7.079	Night
-7.057	Every	-7.117	Street	-7.054	Hope
-7.048	Indeed	-7.094	Long	-7.033	Never
-7.043	Mr.	-7.076	Know	-7.009	May
-7.038	Never	-7.070	Strange	-6.995	Shall
-7.034	Length	-7.062	Many	-6.987	Ever
-7.031	Two	-7.057	Come	-6.945	Death
-6.992	Still	-7.049	Even	-6.945	Towards
-6.971	Long	-7.014	Knew	-6.941	Adrian
-6.962	Must	-6.998	Great	-6.893	First

Table 8

	SVM TF-IDF, POS, LDA			SVM TF-IDF			
	Precision	Recall	F1	Precision	Recall	F1	Support
0	0.82	0.84	0.83	0.80	0.82	0.81	1570
1	0.82	0.80	0.81	0.81	0.79	0.80	1071
2	0.83	0.82	0.82	0.81	0.80	0.81	1275
Accuracy			0.82			0.81	
Macro Average	0.82	0.82	0.82	0.81	0.81	0.81	3916
Weighted Average	0.82	0.82	0.82	0.81	0.81	0.81	3916

Table 9

	SVM TF-IDF, POS, LDA			SVM TF-IDF			
Actual	EAP	MWS	HPL	EAP	MWS	HPL	Totals
EAP	1323	116	131	1293	126	151	1570
MWS	132	859	80	139	846	86	1071
HPL	163	71	1041	177	74	1024	1275
Totals	1618	1046	1234	1609	1046	1261	3916

Table 10

	SVM POS, LDA, Func. Words			
	Precision	Recall	F1	Support
0	0.68	0.77	0.72	1570
1	0.69	0.64	0.66	1071
2	0.71	0.63	0.67	1275
Accuracy			0.69	3916
Macro Average	0.69	0.68	0.68	3916
Weighted Average	0.69	0.69	0.69	3916

Table 11

	SVM POS, LDA, Func. Words			
Actual	EAP	MWS	HPL	Totals
EAP	1206	162	202	1570
MWS	256	682	133	1071
HPL	323	149	803	1275
Totals	1785	993	1138	3916

Table 12

Edgar Allen Poe		Mary Shelley		HP Lovecraft	
1.513	Kate	1.583	Tense	1.719	Agonizing
1.514	Endeavor	1.611	Romero	1.742	Sister
1.517	Waned	1.615	Unnamable	1.744	Misery
1.521	Invention	1.620	Musides	1.745	Krempe
1.522	Usher	1.627	Weird	1.776	Protector
1.527	Nose	1.645	Aout	1.778	Waves
1.536	Altogether	1.648	Warren	1.778	Justine
1.538	Legs	1.650	Uncle	1.813	Pride
1.538	Honor	1.663	Git	1.829	Going far
1.556	Ordinary	1.678	Outside	1.837	Winter

3.2 Support Vector Machine

Deploying the SVM is aimed to achieve the same task as the MNB classifier and is trained and evaluated using the same data partitions so that we may compare results and performance. We can see that the SVM performance for both sets of features are quite comparable to the best combination of MNB features: TF-IDF features only. We can see that the performance of both SVM models are better than either of the MNB models from the performance metrics recorded in Table 8. We can see that the feature set which incorporates the style and structure features performs better than the text-only feature set, but only marginally. However, we can also see the top features for this feature set (TF-IDF, POS & LDA) in Table 12, with the surprising characteristic which illuminates support against the literature's contemporary consensus on authorship classification where a model which incorporates both text content features and a combination of style and structure features, the model appears to be utilizing the text features more-so than the style, structure, and semantic features to make a prediction. As a validation step to see if the authorship attribution can be performed better utilizing only structure, style, and function word features, just as we explored with MNB, we attempt to see if performance improves under these conditions utilizing SVM. We can observe that the performance does not improve - just as we observed with MNB - using only these features and is further evidence contrary to the contemporary consensus on predictors of authorship.

4 Discussion

With the performance of authorship classification being achieved with relatively good performance, we can draw some insight on the comparison of feature sets. We observe that SVM outperforms MNB, and more specifically, we find that the SVM model trained on a combination of content features, style features and structure features out-performs an SVM model which utilizes only the content features. While authorship attribution has come to a consensus that this is to be expected, observing the most predictive features shows us that it was in fact the content features that provided most of the predictive support for any given class. In conjunction with our MNB analysis, we can conclude that within the scope of this classification task, style and structure characteristics do not yield a better authorship classification – contrary to literature consensus [1] [2] [3].

As an additional step, we observed the addition of function words - another style characteristic - and removed the content features, which are typically seen as noise but have yielded the best predictions in our case, and we observe a stark decline in authorship attribution performance across both models. These results point to the fact that while style and structure can improve authorship attribution classification, these characteristics may not be the best predictors of authorship. One caveat, discussed early in this paper, is the structure of the data, being a corpus consisting of sentences as opposed to a document of multiple sentences. This is different than typical authorship attribution tasks, however, further work on this matter within the context of a different corpus may show that sentences may be the proper lens within which authorship should be observed.

References

[1] D. Holmes and R. Forsyth, "The Federalist revisited: New directions in authorship attribution," *Literary and Linguistic computing,* pp. 111-127, 1995.

[2] P. Juola, "Authorship attribution," *Foundations and Trends in Information Retrieval,* pp. 233-334, 2008.

[3] J. O'shea, Z. Bandar and K. Crockett, "A Multi-classifier Approach to Dialogue Act Classification Using Function Words," *Transactions on Computational Collective Intelligence VII,* pp. 119-143, 2012.

[4] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[5] M. Hoffman, F. Bach and D. Blei, "Online Learning for Latent Dirichlet Allocation," *Advancements in Neural Information Processing Systems 23,* 2010.

[6] M. Röder, A. Both and H. Alexander, "Exploring the space of topic coherence measures," *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015.

[7] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63-70, 2014.

[8] J. O'Shea, "Semantic Similarity Function Words Lists," [Online]. Available: https://semanticsimilarity.wordpress.com/function-word-lists/. [Accessed 09 05 2021].