Topic or Style? Exploring the Most Useful Features for Authorship Attribution

Yunita Sari, Mark Stevenson and **Andreas Vlachos**
Department of Computer Science
University of Sheffield, UK
{y.sari, mark.stevenson, a.vlachos}@sheffield.ac.uk

Abstract

Approaches to authorship attribution, the task of identifying the author of a document, are based on analysis of individuals' writing style and/or preferred topics. Although the problem has been widely explored, no previous studies have analysed the relationship between dataset characteristics and effectiveness of different types of features. This study carries out an analysis of four widely used datasets to explore how different types of features affect authorship attribution accuracy under varying conditions. The results of the analysis are applied to authorship attribution models based on both discrete and continuous representations. We apply the conclusions from our analysis to an extension of an existing approach to authorship attribution and outperform the prior state-of-the-art on two out of the four datasets used.

1 Introduction

Authorship attribution plays an important role in many applications, including plagiarism detection and forensic investigation. Approaches to this problem attempt to identify a document's author through analysis of individual's writing style and/or topics they tend to write about. The problem has been extensively studied and a wide range of features has been explored (Stamatatos, 2013; Schwartz et al., 2013; Seroussi et al., 2013; Hürlimann et al., 2015). However, there has been a lack of analysis of the behavior of features across multiple datasets or using a range of classifiers. Consequently, it is difficult to determine which types of features will be most useful for a particular authorship attribution dataset.

Authorship attribution is a unique task which is closely related to both the representation of individuals' writing style and text categorization. In some cases, where there is a clear topical distinction between the documents written by different authors, content-related features such as those used in text categorization may be effective. However, style-based features are more likely to be effective for datasets containing a more homogeneous set of topics. Previous work on feature exploration for authorship attribution, focused on the overall effectiveness of features without considering the characteristics of the datasets to which they were applied, e.g. (Grieve, 2007; Guthrie, 2008; Stamatatos, 2009; Brennan et al., 2012; Sapkota et al., 2015). A wide range of features have been applied to the authorship attribution problem and many previous studies concluded that using character n-grams is often effective, e.g. (Peng et al., 2003; Koppel et al., 2011; Schwartz et al., 2013; Sapkota et al., 2015; Sari et al., 2017; Shrestha et al., 2017). Thus, character n-grams have become the *go-to* features for this task to capture both an author's topical preferences and writing style.

This study explores how the characteristics of a dataset affect the usefulness of different types of features for the authorship attribution task. Experiments are carried out using four datasets that have previously been widely used for this task. Three types of features are considered: *style*, *content* and *hybrid* (a mixture of the previous two types). In contrast to previous work, this study finds that character n-grams do not perform equally well in all datasets. The analysis holds for authorship attribution models using discrete and continuous representations. Using topic modeling and feature analysis, the most effective features can be successfully predicted for three of the four datasets. The results of this analysis

Proceedings of the 27th International Conference on Computational Linguistics, pages 343–353
Santa Fe, New Mexico, USA, August 20-26, 2018.

are applied via a novel extension of a recently proposed neural approach (Sari et al., 2017) and improved state-of-the-art performance are obtained for two of the four datasets.

2 Related Work

Authorship attribution features are often referred to *stylometric features* since the main goal of the task is often thought to be modeling the authors' writing style. Grieve (2007) conducted experiments which involved thirty-nine different types of textual measurements commonly used in attribution studies. His experiments, which were performed using the Chi-squared test on The Telegraph Columnist corpus, concluded that the combination of word and punctuation mark profiles are effective features for representing authors. Similar to Grieve, Guthrie (2008) carried out an exploration of 166 features used for authorship attribution including commonly used stylistic features and several others intended to capture emotional tone. He reported that fifteen features, including punctuation marks, pronouns, fog index and average sentence length to be the most useful. Stamatatos (2009) divided authorship attribution features into five groups: lexical, character, syntactic, semantic and application-specific features. Compared to others, lexical and character features are commonly used in authorship attribution work as they provide rich information about the author's topical preferences and writing style. In addition, both types of features can be extracted in many languages and datasets with little effort.

Simple lexical features (e.g. word frequencies, word n-grams, function words, hapax legomena, word/sentence length) have been widely used since early attribution work (Mendenhall, 1887). Function words have been proved to be effective features and several studies have reported their usefulness, e.g. (Mosteller and Wallace, 1964; Argamon and Levitan, 2005; Koppel et al., 2005; Juola and Baayen, 2005; Zhao and Zobel, 2005). Bag-of-words approaches have also been reported as being useful for authorship attribution (Koppel et al., 2011). These approaches are also commonly applied for sentiment analysis and topic classification tasks (Zhang et al., 2015; Heap et al., 2017).

The usefulness of character n-grams has been highlighted in several studies including (Peng et al., 2003; Stamatatos, 2013; Schwartz et al., 2013; Sapkota et al., 2015). Koppel et al. (2011) argued that this effectiveness comes from their ability to capture both content and stylistic information. Similar conclusions were reported by Sapkota et al. (2015). They analysed subgroups of character n-grams in both single and cross-domain settings. They concluded that affixes and punctuation n-grams make a significant contribution towards the effectiveness of character n-grams.

Our study differs from previous work in that we perform dataset analysis using topic modeling followed by feature ablation experiments. Thus, we are able to determine the level that each type of feature affects authorship attribution accuracy.

3 Datasets

Experiments[1] are performed using four authorship attribution datasets: Judgment (Seroussi et al., 2011), CCAT10, CCAT50 (Stamatatos, 2008), and IMDb62 (Seroussi et al., 2010). These datasets were chosen because they are all commonly used in previous literature and represent a range of characteristics in terms of the number of authors, topic/genre and document length (see Table 1).

	Judgment	CCAT10	CCAT50	IMDb62
genre	legal judgments	newswire		movie reviews
# authors	3	10	50	62
# total documents	1,342	1,000	5,000	79,550
avg characters per document	11,957	3,089	3,058	1,401
avg words per document	2367	580	584	288

Table 1: Dataset statistics.

[1]Code to reproduce the experiments is available from `https://github.com/yunitata/coling2018`

Judgment consists of legal judgments from three Australian High Court judges while both CCAT datasets are subsets of Reuters Corpus Volume 1 (RCV1) (Rose et al., 2002). The IMDb62 dataset was collected from movie reviews and message board posts of the Internet Movie database. Train/test partitions are provided for both CCAT datasets by the respective authors. For Judgment and IMDb62 we follow previous work (Seroussi et al., 2013) by using 10-fold cross validation in our experiments. We do not make use of datasets from recent authorship attribution shared task events, e.g. PAN (Juola, 2012), due to their relatively small size and fact that they provide a very small number of documents per author.

4 Dataset Analysis

The aim of this analysis is to quantify the topical similarity between authors in each of the datasets considered. The motivation for this is that certain datasets may have clear topical preferences between authors which cause authorship attribution to be biased towards topic classification. Therefore, topic modeling can help assess the topical (dis-)similarity among authors. We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative probabilistic model for text collections. Documents are represented as mixtures over latent topics, where each topic is characterized by a distribution over words. Assuming a trained topic model over an authorship attribution dataset D, if C_α is the set of documents written by author α and Φ_i is the topic distribution for the i-th document in C_α, then we estimate the topic distribution for a particular author, θ_α, as follows:

$$\theta_\alpha = \frac{\sum_{i=1}^{|C_\alpha|} \Phi_i}{|C_\alpha|} \tag{1}$$

Following this, the difference between two author's topic probability distributions θ_α and θ_β is calculated using the Jensen-Shannon Divergence (JSD) (Cover and Thomas, 2006):

$$JSD(\theta_\alpha, \theta_\beta) = \frac{1}{2} D_{KL}(\theta_\alpha || M) + \frac{1}{2} D_{KL}(\theta_\beta || M) \tag{2}$$

where $M = \frac{1}{2}(\theta_\alpha + \theta_\beta)$ and D_{KL} is Kullback-Leibler divergence.

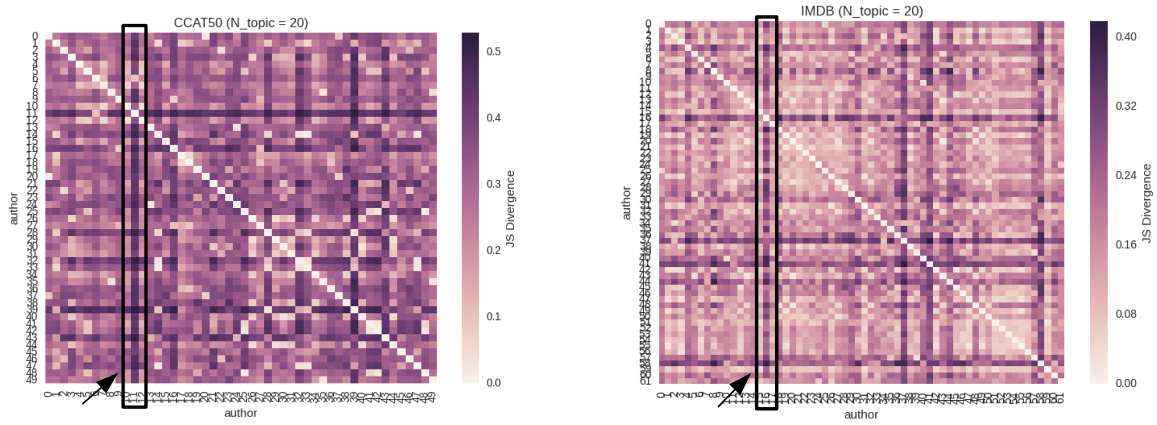
Table 2 shows the average of JSD for all author pairs in each of the datasets having trained a topic model with varying numbers of topics. High JSD scores indicate more topical dis-similarity between authors in the dataset.[2] The CCAT datasets, which contain on-line news, have higher scores compared to Judgment and IMDb62. The scores for CCAT50 and CCAT10 are similar, despite the fact that the first dataset contains five times the number of authors of the second. The consistency of this comparison across different numbers of topics indicates that this method of assessing content similarity between authors is robust with respect to tuning this parameter. Judgment has the lowest score across the four datasets indicating that the authors discuss the most similar topics. Finally, scores for the IMDb62 dataset obtained were higher than those for Judgment but lower than both CCAT's scores. Differences in scores for IMDb62 are due to the authors' preferences, as some comment on the story while others comment on the characters of the movie. Overall, we observe that the genre of the datasets influences the topical dis-similarity between authors.

Confusion matrices were created to further analyse the differences between authors. These matrices were generated after running LDA with 20 topics for 1000 iterations. Similar patterns were observed using different numbers of topics. Darker color indicates higher JSD score between two authors. In the CCAT50 dataset (Figure 1a), one author (number 11, indicated by an arrow) has very different topic preferences compared to the others. Articles written by author 11 mainly discuss topics related to *gold, exploration, Canada, Indonesia* which are rarely picked by the other authors. A similar pattern is found in IMDb62 (see Figure 1b), where reviews by author 16 (also indicated by an arrow) are dominated by positive comments about movies unlike other authors who tended to write negative reviews or discuss the story and/or the characters. However, unlike the aforementioned datasets, authors in Judgment wrote about relatively similar topics.

[2] We do not assume linear scaling.

#topic	Judgment	CCAT10	CCAT50	IMDb62
3	0.0056	0.2053	0.1785	0.1000
10	0.0148	0.3010	0.2867	0.1471
20	0.0180	0.3193	0.3279	0.1617
30	0.0256	0.3414	0.3269	0.1627
40	0.0272	0.3417	0.3291	0.1681
50	0.0281	0.3403	0.3326	0.1634

Table 2: Average JS Divergence



(a) CCAT50 (train)

(b) IMDb62

Figure 1: Author topic distribution (20 topics).

5 Feature Analysis

An ablation study was carried out to determine the contribution of different features for each dataset. Following previous studies (Abbasi and Chen, 2008; Stamatatos, 2009), feature groups are divided into three types (see Table 3):

- **Style:** Style-based features, such as usage of function words, digits and punctuation, capture an author's writing style. We used pre-defined sets of 174 function words and 12 punctuation marks.
- **Content:** Content-based features, e.g. bags of n-grams, represent the author's topical preferences. All function words are removed when extracting these features.
- **Hybrid:** The final feature type, hybrid, are character n-grams which are intended to capture both writing style and topical preferences (Koppel et al., 2011; Sapkota et al., 2015).

Both character and word *n*-grams are limited to bi- and tri-grams. As the purpose of these ablation experiments is not to outperform previous work, only the 100 most common n-grams are used for each feature type.

Authorship attribution experiments were carried out using two classifiers: a single hidden layer Feed-forward Neural Network model (FNN) and Logistic Regression (LR). The FNN hyper-parameters including the number of neurons and dropout rates were tuned on the development set for each of the datasets. For Judgment, CCAT10 and CCAT50, we set the number of epochs to 250 and 100 for IMDb62. For all datasets, early stopping was used on the development sets and the models were optimized with the Adam update rule (Kingma and Ba, 2015). Since none of the datasets have a standard development set, we randomly selected 10% of the training data for this purpose. For LR, we found that using the default parameters from Scikit-Learn (Pedregosa et al., 2011) resulted in comparable performances to the FNN. Accuracy was used as the evaluation metric to measure authorship attribution performance.

Type	Group	Category	#	Description
Style	Lexical	Word-level	2	Average word length, number of short words
		Char-level	2	Percentage of digits, percentage of uppercase letters
		Letters	26	Letter frequency
		Digits	10	Digit frequency
		Vocabulary richness	2	Richness (hapax-legomena and dislegomena)
	Syntactic	Function words	174	Frequency of function words
		Punctuation	12	Occurrence of punctuation
Content	Word *n*-gram	Words unigrams	100	Frequency of 100 most common word unigrams
		Words bigrams	100	Frequency of 100 most common word bigrams
		Word trigrams	100	Frequency of 100 most common word trigrams
Hybrid	Char *n*-gram	Char bigrams	100	Frequency of 100 most common character bigrams
		Char trigrams	100	Frequency of 100 most common character trigrams

Table 3: Authorship attribution feature sets.

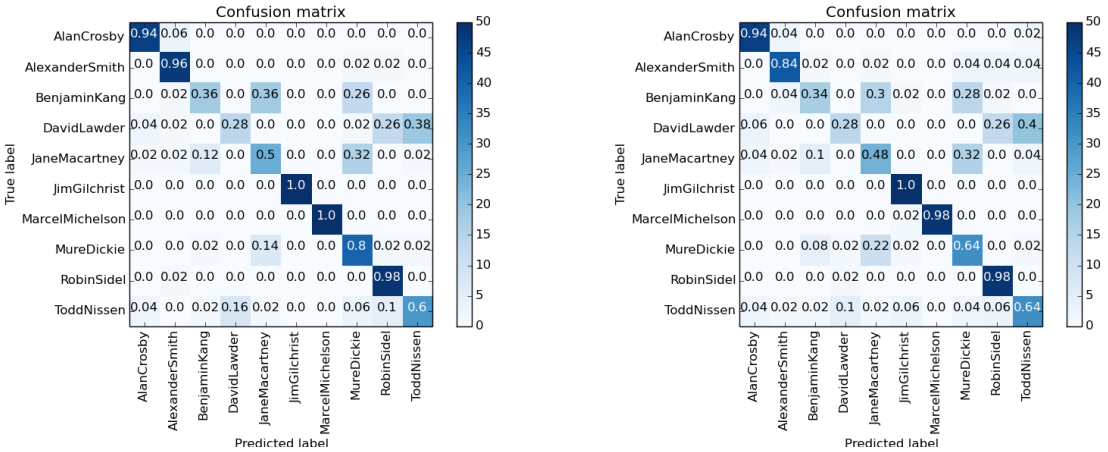
5.1 Feature Ablation Results

Results are presented in Table 4. The $(-)$ symbol indicates that the respective feature type is excluded. The results confirm our topic model-based analysis (see Section 4). Style-based features are more effective for datasets in which authors discuss similar topics, e.g. Judgment and IMDb62. As expected, content-based features are generally more effective when there is more dis-similarity between the topics discussed by the authors in the dataset, e.g. CCAT10 and CCAT50, but are of limited usefulness when the topics are similar (particularly for the Judgment dataset). The hybrid features appear to behave similarly to the content-based features since they are most useful when the topic dis-similarity between authors is high.

Features	Judgment		CCAT10		CCAT50		IMDb62	
	FNN	LR	FNN	LR	FNN	LR	FNN	LR
all features	89.43	90.02	75.40	74.20	60.20	60.56	85.25	85.00
(–) Style	**-3.87**	**-4.32**	-3.00	+0.40	-3.40	-2.60	**-6.91**	**-8.39**
(–) Content	-1.43	+0.30	**-3.60**	**-3.00**	**-4.52**	-4.08	-2.77	-2.68
(–) Hybrid	-0.83	-0.29	-3.40	-1.00	-1.28	**-4.68**	-2.02	-5.32

Table 4: Feature ablation results.

To examine the results further, we generated confusion matrices for the Logistic Regression (LR) classifier applied on CCAT10 dataset (Figure 2). The effect of removing content-based features is shown in Figure 2b where the prediction accuracy for authors *Alexander Smith* and *Mure Dickie* drops from 96% and 80% (see Figure 2a) to 84% and 64% respectively. Content-based features are essential in this particular genre (newswire) dataset, since each author usually has different topical interests. For example, among the ten authors in the dataset, *Alexander Smith* mostly discussed topic related to *investment and*



(a) All features (b) (−) Content

Figure 2: Confusion Matrices of LR classifier with different features types on CCAT10.

finance, while *China* was dominantly written about by *Mure Dickie, Benjamin Kang* and *Jane Macartney*. In addition, writing style between authors in this genre can be very similar. Thus, applying style-based or hybrid features alone may not be effective.

Additional feature exploration was carried out to analyse what types of features are more important to the classifier overall. We performed an analysis using LIME (Ribeiro et al., 2016), a model agnostic framework for intepretability. LIME provides explanations about how a classifier made a prediction by identifying important input features. We selected a document from each of the datasets and analysed what kind of features were used in its prediction. Figures 3, 4, 5 and 6 present the predictions of LR trained on 1000 word unigrams in Judgment, CCAT10, CCAT50, and IMDb62 respectively. In these experiment function words are not removed. For each of the documents presented, LR made correct author predictions with probability close to 100%. The darker shade indicates more important words in the attribution prediction. In the two CCAT datasets the classifier put more weight on content-based words such as *Thomson, Canada* and *Toronto*. In contrast, function words e.g. *at, had, and, was* appear to be more salient in Judgment and IMDb62.

.) . but in bracton 's account of the great convention at merton we read only of those qui nati and not those qui geniti , fuerunt ante sponsalia vel matrimonium , (fol . :) and we learn elsewhere from him (fol .) that it mattered not for legitimacy whether the child was begotten and born after the marriage or begotten before but born into the marriage or begotten in the marriage and born after the marriage had been dissolved and whether dissolved by death or by divorce and that it did not matter whether the union was by matrimonium or (subject to certain exceptions) by sponsalia . but we find that he recognizes adulterine bastardy . the child is to be presumed a bastard , it appears , if feebleness , frigidity or impotence of the husband is proved per multum tempus or absence for two years from the kingdom or from the county or shire is shown . if he returns and finds his wife pregnant or that she has a child of a year or less , whether he avows it and nurtures it or not , the child may rightly be excluded from succession because it could not be heir . but , if it was possible to presume that he could have engendered the child , it seems that decisive importance was given to avowal and nurture by

Figure 3: Important word unigrams features in Judgment.

Alcatel Alsthom said on Monday it was in talks with Aerospatiale and Dassault about a joint offer for the government's 58-percent stake in defence electronics group Thomson-CSF.
Prime Minister Alain Juppe said a decision about the procedure for the privatisation of Thomson-CSF would be made before the end of February.
"There are discussions with the companies mentioned in the press," an Alcatel spokesman said when asked to react to newspaper reports about a joint bid.
Industry sources said that Alcatel chairman Serge Tchuruk had kept the government informed about his plans to form an alliance with Aerospatiale and Dassault in order to win the Thomson-CSF stake.
Alcatel in October lost out to Lagardere Groupe in bidding for state-controlled Thomson SA, which has the stake in Thomson-CSF as well as 100 percent of consumer electronics group Thomson Multimedia (TMM).
But the government had to suspend the sale on December 4 after the independent Privatisation Commission balked against the terms of the sale by Lagardere of TMM to Daewoo Electronics of South Korea.

Figure 4: Important word unigrams features in CCAT10.

We also observed a document in the IMDb62 dataset where the classifier assigns similar prediction

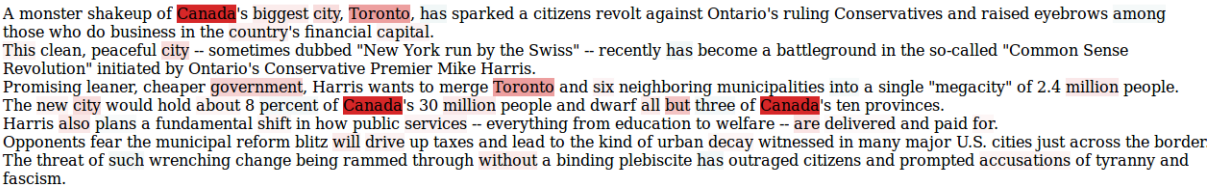


Figure 5: Important word unigrams features in CCAT50.

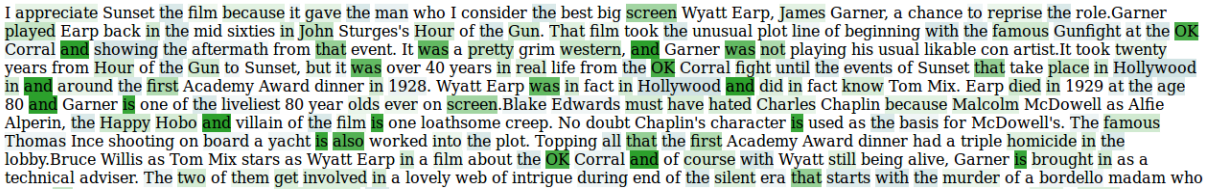


Figure 6: Important word unigrams features in IMDb62.

probabilities to two authors as presented in Figure 7. The classifier put the same weight to function words *and* and *to* which represent two different classes of authors, 26 and not 26 (the LR classifier uses a one-versus-all scheme). The correct decision of the classifier is more likely helped by the presence of some less significant features such as *is, becomes, There, usual* and *could*.

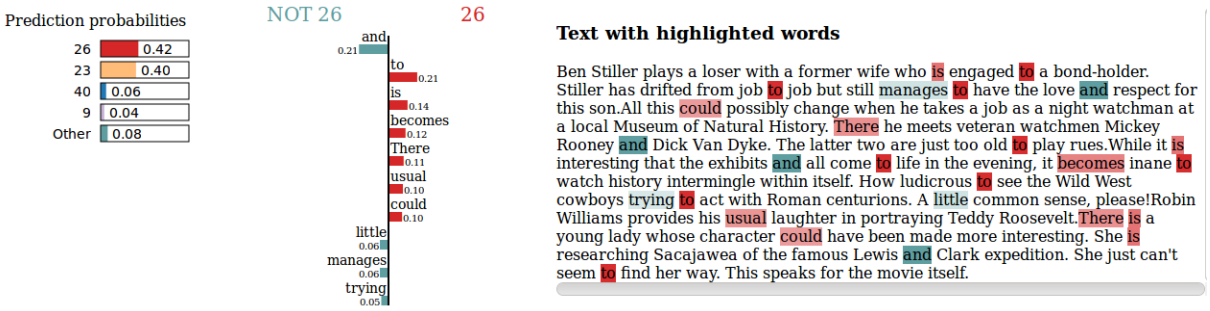


Figure 7: Explanation of individual predictions of Logistic Regression classifier on an IMDb62 document using LIME. The bar chart represent the weight given to the most relevant words which are also highlighted in the text.

6 Extending a Neural Model

These findings are further validated by applying them to a continuous n-grams-based authorship attribution model recently proposed by Sari et al. (2017). They represented a document as a bag of n-grams features and learned the continuous representation of each feature jointly with the classifier in a shallow feed-forward neural network (Joulin et al., 2017). Sari et al. conducted experiments with three different feature choices: characters, words and their combination. The character-based model outperformed the state-of-the-art on the CCAT50 and IMDb62 datasets, while producing comparable results on the remaining two.

We extend their character-based model by incorporating each feature type (*style, content* and *hybrid*) as an auxiliary feature represented in discrete form. Auxiliary features provide additional information related to the dataset characteristics. Given x_{aux} as a normalized auxiliary features frequency vector, V is the weight applied to the features and f is the activation function (ReLu), the hidden layer h performs the following computation:

$$h = f(Vx_{aux}) \tag{3}$$

The probability distribution over the label for a document then can be described as:

$$p(y|x) = softmax(W_{out}[Ax, h]) \tag{4}$$

where x is the frequency vector of features for the document, A is the embedding matrix, W_{out} is the weight matrix of the output layer and $[Ax, h]$ is the concatenation of Ax and h.

All the character n-gram embeddings and hidden layer in the model were initialized using Glorot uniform initialization (Glorot and Bengio, 2010). We used the best hyper-parameters values for each of the datasets which have been tuned in the development set via a small grid search over all combinations of embedding size and dropout rate (specifically dropout in the concatenation layer). The hidden size of hidden auxiliary layer was set to 2. For the rest of the hyper-parameters, we used the values from the baseline model (Sari et al., 2017). For Judgment, CCAT10 and CCAT50, we set the number of epochs to 250 and 100 for IMDb62. For all datasets, early stopping was used on the development sets and the models were optimized with the Adam update rule (Kingma and Ba, 2015).

6.1 Results

Table 5 presents the results of the experiment and compares them against previously reported ones on the same data sets. In the bottom portion of the table it can be seen that for each of the four data sets there is at least one feature type which leads to improved results when it is incorporated into the model. Our results demonstrate that better performance can be achieved by taking the data characteristic into account on choosing authorship attribution features. Moreover, the results provide evidence that character n-grams which have been known as typical *go-to* features do not perform equally well in all types of datasets. For three datasets (CCAT10, CCAT50 and IMDb62) the best result is obtained using the feature type identified as being most useful in Section 5. However, we find that using the style features does not improve results on the Judgment dataset as we had expected. The relatively poor performance of the style features may be due to the baseline model (the continuous character n-grams) which effectively captured all the author's writing style. Thus the addition of auxiliary style features did not lead to any improvement.

The results reported here for the CCAT50 and IMDb62 datasets outperform the previously best reported results (Sari et al., 2017) and the model reported here therefore represents a new state-of-the-art performance. The improvements for IMDb62 are statistically significant ($p < 0.05$, paired t-test).

7 Conclusions

This paper describes experiments on the relationship between the effectiveness of different types of features for authorship attribution and characteristics of datasets. We find that the most effective features for datasets can be predicted by applying topic modeling and feature analysis. Content-based features tend to be suitable for datasets with high topical diversity such as the one constructed from on-line news. Datasets with less topical variance, e.g. legal judgments and movie reviews, benefit more from style-based features. The effectiveness of our proposed analysis is further validated by the performance of our proposed neural model which achieved the state-of-the-art results on two datasets.

Model	Judgment	CCAT10	CCAT50	IMDb62
Previous work				
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	69.30	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	80.80	-	-
STM-Asymmetric cross (Plakias and Stamatatos, 2008)	-	78.00	-	-
SVM with bag of local histogram (Escalante et al., 2011)	-	**86.40**	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	92.52
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79
Baseline model				
Continuous character n-gram (Sari et al., 2017)	91.29	74.80	72.60	94.80
Proposed model				
(+) style	91.07	76.00	72.72	**95.93***
(+) content	91.51	76.20	**72.88**	95.59
(+) hybrid	91.21	74.80	71.76	95.26

Table 5: Results with comparison against baseline and previous work.
* denotes significant improvement over baseline model ($p < 0.05$).

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. The first author would like to acknowledge Indonesia Endowment Fund for Education (LPDP) for support in the form of a doctoral studentship.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2).

Shlomo Argamon and Shlomo Levitan. 2005. Measuring the Usefulness of Function Words for Authorship Attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security*, 15(3):1–22, November.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local Histograms of Character N-grams for Authorship Attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 288–298, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding The Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, AISTATS 2010*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. Society for Artificial Intelligence and Statistics.

Jack. Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270, May.

David Guthrie. 2008. *Unsupervised Detection of Anomalous Text*. Ph.D. thesis, University of Sheffield.

Bradford Heap, Michael Bain, Wayne Wobcke, Alfred Krzywicki, and Susanne Schmeidl. 2017. Word vector enrichment of low frequency words in the bag-of-words model for short text multi-class classification problems. *CoRR*, abs/1709.05778.

Manuela Hürlimann, Benno Weck, Esther van den Berg, Simon Šuster, and Malvina Nissim. 2015. GLAD: Groningen Lightweight Authorship Detection—Notebook for PAN at CLEF 2015. In Linda Cappellato, Nicola Ferro, Gareth Jones, and Eric San Juan, editors, *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org, September.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Patrick Juola and RH Baayen. 2005. A Controlled-corpus Experiment in Authorship Identification by Cross-entropy. *Literary and Linguistic Computing*, pages 1–10.

Patrick Juola. 2012. An Overview of the Traditional Authorship Attribution Subtask. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy*, September.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceeding of the 3rd International Conference for Learning Representations, ICLR 2015*, San Diego, CA, May.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically Determining An Anonymous Author's Native Language. *Intelligence and Security Informatics*, pages 209–217.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship Attribution in The Wild. *Language Resources and Evaluation*, 45(1):83–94, March.

T.C. Mendenhall. 1887. The Characteristic Curves of Composition. *Science*, IX:37–49.

F Mosteller and D. L Wallace. 1964. Inference and Disputed Authorship: The Federalist. Addison Wesley.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fuchun Peng, Dale Schuurmanst, Vlado Kesel, and Shaojun Wan. 2003. Language Independent Authorship Attribution using Character Level Language Models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, Budapest, Hungary.

Spyridon Plakias and Efstathios Stamatatos. 2008. Tensor Space Models for Authorship Identification. In *Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications*, SETN '08, pages 239–249, Berlin, Heidelberg. Springer-Verlag.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June. Association for Computational Linguistics.

T. Rose, M. Stevenson, and M. Whitehead. 2002. The Reuters Corpus - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02)*, pages 827–832, Las Palmas, Canary Islands.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not All Character N-grams Are Created Equal: A Study in Authorship Attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–102, Denver, Colorado, May–June. Association for Computational Linguistics.

Yunita Sari, Andreas Vlachos, and Mark Stevenson. 2017. Continuous N-gram Representations for Authorship Attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia, Spain, April. Association for Computational Linguistics.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship Attribution of Micro-Messages. In *Proceeding of Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891, Seattle, USA.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert, 2010. *Collaborative Inference of Sentiments from Texts*, pages 195–206. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yanir Seroussi, Russell Smyth, and Ingrid Zukerman. 2011. Ghosts from the High Courts past: Evidence from computational linguistics for Dixon ghosting for McTiernan and Rich. *University of New South Wales Law Journal*, 34(3):984–1005.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2013. Authorship Attribution with Topic Models. *Journal Computational Linguistics*, 40(2):269–310.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain, April. Association for Computational Linguistics.

Efstathios Stamatatos. 2008. Author identification: Using Text Sampling to Handle The Class Imbalance Problem. *Information Processing and Management*, 44(2):790 – 799.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March.

Efstathios Stamatatos. 2013. On The Robustness of Authorship Attribution based on Character n-gram Features. *Journal of Law and Policy*, 21(2):421–439.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Montreal, Canada. MIT Press.

Ying Zhao and Justin Zobel. 2005. Effective and Scalable Authorship Attribution Using Function Words. *Information Retrieval Technology*, pages 174–189.