Abstract

Authorship attribution is the problem of identifying the most plausible author of an anonymous text from a set of candidate authors. Researchers have investigated same-topic and cross-topic scenarios of authorship attribution, which differ according to whether unseen topics are used in the testing phase. However, neither scenario allows us to explain whether errors are caused by failure to capture authorship style, by the topic shift or by other factors. Motivated by this, we propose the topic confusion task, where we switch the author-topic configuration between training and testing set. This setup allows us to probe errors in the attribution process. We investigate the accuracy and two error measures: one caused by the models’ confusion by the switch because the features capture the topics, and one caused by the features’ inability to capture the writing styles, leading to weaker models. By evaluating different features, we show that stylometric features with part-of-speech tags are less susceptible to topic variations and can increase the accuracy of the attribution process. We further show that combining them with word-level n-grams can outperform the state-of-the-art technique in the cross-topic scenario. Finally, we show that pretrained language models such as BERT and RoBERTa perform poorly on this task, and are outperformed by simple n-gram features.

1 Introduction

Authorship attribution is the problem of identifying the most plausible author of an anonymous text from a set of candidate authors. This problem is important as it can reveal characteristics of an author given a relatively small number of their writing samples. Authorship attribution has been applied to various domains, the first of which is literature (Mosteller and Wallace, 1963; Juola, 2008), and currently it is being used in criminal investigations where a domain-expert would use authorship techniques to help law enforcement identify the most plausible author of an anonymous, threatening text (Ding et al., 2015; Rocha et al., 2016).

Previous authorship attribution evaluations can be classified into same-topic or cross-topic, depending on whether new, unseen, topics are introduced at test time. Figure 1a shows the same-topic scenario where all pairs of authors and topics in the test set are attested at training time. This setup, however, is considered artificial (Koppel et al., 2009) as it is unlikely that writing samples of candidate authors across multiple topics would be available in real-life applications.

The cross-topic scenario (Stamatatos, 2013) is a more realistic one where new, unseen topics are introduced in the test phase, as shown in Figure 1b. This task requires attribution techniques to focus on cues that capture the authors’ writing style across topics. Compared to the same-topic scenario, performance of well-known authorship methods using word- and character-level n-gram features is much lower, as they relied on topic-specific cues to infer authorship.

Both, same- and cross-topic scenarios make strong assumptions about the relationship between topics and the writing style. Same-topic scenarios naively assume the availability of writing samples on all the topics by all authors at training time. Cross-topic scenarios account for the topics effect, but consider them independent from the author. This setup fixes the topics in training and testing, which sheds little light on the authorship attribution errors—in particular whether these errors are caused by the topic change or by the model’s inability to capture the authors’ writing styles. Additionally, this setup assumes that the investigated document will always be on a new, unseen topic. The following example can explain why the latter assumption is dangerous. A criminal investigating might have a candidate author who had produced a threatening message in their past which will almost certainly be used in training. Given that attribution techniques are highly influenced by the topic, that person might be convicted based on the topic.
similarity with the investigated text.

Traditionally, the evaluation of new methods or features that are introduced to address the cross-topic issue has been based on the difference in the accuracy either on the attribution process, or on ablation studies. While this methodology enhanced the performance on the downstream task and helped answer which features perform well, there is a need for methods that can help us understand why certain features are performing better.

In this work, we propose a new evaluation setting, namely, the topic confusion task. We propose to control the topic by making it dependant on the author, and switching the topic-author pairs between training and testing, as shown in Figure 1c. The setup allows us to measure the degree to which certain features are influenced by the topic, as opposed to the author’s identity. The intuition is that, the more a feature is influenced by the topic of the document to identify its most plausible author, the more confusing it will be to the classifier when the topic-author combination is switched, which will lead to worse authorship attribution performance. To better understand the writing style and the capacity of the used features, we use the accuracy and split the error on this task to one that is caused by the models’ confusion by the topics, and one caused by the features’ inability to capture the authors’ writing styles.

The primary contributions of this work are summarized as follows:

• We propose topic confusion as a new scenario in authorship attribution, and use it to measure the effectiveness of features in the attribution process.

• Our results show that word-level n-grams can easily outperform pretrained embeddings from BERT and RoBERTa models when used as features for cross-topic authorship attribution. The results also show that a combination of n-grams on the Part-Of-Speech (POS) tags and stylistic features, which were outperformed by word- and character-level n-grams in earlier works on authorship attribution can indeed enhance cross-topic authorship attribution. Finally, when these features are combined with the current state of the art, we achieve a new, higher accuracy.

• We present a cleaner, curated, and more balanced version of the Guardian dataset to be used for future work on both same-topic, and cross-topic authorship attribution. The main goal is to prevent any external factors, such as the dataset imbalance, from affecting the attribution results.

2 Related Work

2.1 Same-Topic Authorship Attribution

Early approaches to authorship attribution depended on manual inspection of the textual documents to identify the authors’ writing patterns. Mendenhall (1887) showed that word lengths and frequencies are distinct among authors. This, however, was a tedious task due to having large amounts of text to analyze. The first work that used a computational approach is (Mosteller and Wallace, 1963), which used the Naïve Bayes algorithm with the frequency of function words to identify the authors of the Federalist papers (Juola, 2008; Koppel et al., 2009; Stamatatos, 2009).

Research efforts have aimed at finding new sets of features for current domains/languages, adapting existing features to new languages or communication domains, or using new classification techniques (Kešelj et al., 2003; Abbasi and Chen, 2005, 2006; Frantz, 2007; Koppel et al., 2009; Stamatatos, 2009, 2013, 2017; Silva et al., 2011; Layton et al., 2012; Schwartz et al., 2013; Iqbal et al., 2013; Sidorov et al., 2014; Sanchez-Perez et al., 2017; Zhang et al., 2018). Alternatively, different elements of and constraints on the attribution process have been investigated, motivated by the real-life applications of authorship attribution. (Houvardas and Stamatatos, 2006; Luyckx and Daelemans, 2011; Ding et al., 2015, 2019; Al-takrori et al., 2018).
2.2 Cross-Topic Authorship Attribution

There have been more recent attempts to investigate authorship attribution in more realistic scenarios, and many studies emerged where the constraints differ from the training to the testing samples. Examples of these studies are (Bogdanova and Lazaridou, 2014) on cross-language, (Goldstein et al., 2009; Custódio and Paraboni, 2019) on cross-domain/genre, and finally, (Mikros and Argiri, 2007; Overdorf and Greenstadt, 2016; Sundararajan and Woodard, 2018; Stamatos, 2013, 2017, 2018; Barlas and Stamatos, 2020) on cross-topic which is the focus of this work.

Recently, Stamatos (2017, 2018) achieved state-of-the-art results on cross-topic authorship attribution on the Guardian dataset (Stamatatos, 2013). Their proposed character- and word-level \(n\)-grams approach is motivated by text distortion (Granados et al., 2012) for topic classification. Stamatos (2013) kept the most frequent words and masked the rest of the text.

Datasets for cross-topic authorship attribution are scarce and small in size. This is because these datasets require writing samples on each one of the topics for all the authors. The largest authorship attribution dataset in terms of average number of documents per author-topic is the one provided by Stamatos (2013). It contains a total of 381 articles for 13 authors on four topics. Each author has around 7.3 articles per topic, and approximately 29.3 articles in total.

2.3 Neural Methods for Authorship Attribution

Ruder et al. (2016), Ge et al. (2016), Ding et al. (2019), Shrestha et al. (2017), Sari et al. (2017), and Hitschler et al. (2017) have all shown that their neurally inspired approaches achieve very high accuracy. Still, they require a large amount of data to train from scratch (Zhang et al., 2015) which makes them inapplicable to real-life scenarios with limited data (Luyckx and Daelemans, 2011).

Posadas-Durán et al. (2017) and Gómez-Adorno et al. (2018) managed to train the well-known document-to-vector (doc2vec) architecture proposed by Le and Mikolov (2014) on the Guardian dataset which contains very few samples. We replicated their results on the authorship task and noticed a large drop in performance when the accuracy is weighted relative to the classes weights in the dataset.

Using the weighted accuracy is a common practice to deal with imbalanced datasets and was missing in (Gómez-Adorno et al., 2018).

Barlas and Stamatos (2020) explored the widely used and massively pretrained transformer-based (Vaswani et al., 2017) language models for authorship attribution. Barlas and Stamatos (2020) used the cross-topic dataset in (Goldstein et al., 2009) to fine-tune an RNN-based sequence classification model (Bagnall, 2015) with a pretrained embeddings layer from ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019) and ULMFit (Howard and Ruder, 2018). Although Barlas and Stamatos (2020) outperformed a number of non-neural baselines and achieved an average accuracy of 80.83, they failed to include the results from the same paper (Goldstein et al., 2009) from which they used the dataset. Goldstein et al. (2009) achieved a much higher average accuracy of 94 on the cross-topic scenario by using stop-words and a set of 88 features that include, among others, stylometric features such as average number of words per sentence with Support Vector Machines (SVM) with a nonlinear kernel.

3 The Topic Confusion Task

3.1 Theoretical Motivation

Figure 2a shows the relationship diagram between a document, its author, its topic, and the language rules that govern the writing process\(^1\). According to Ding et al. (2019) these are the factors that affect the process of writing a document. Given a topic’s distribution over words, the author picks a subset of these words and connects them using

\(^1\)There could be other unknown factors that affect any random variable which the attribution process is not aware of.
the language rules which govern what words accompany these topical words and how sentences are structured. Eq. 1 shows the joint probability while ignoring the language model, and assuming the topic distribution is independent from that of the author.

\[
P(A, T, D) = P(A)P(T)P(D|A, T) \tag{1}
\]

\[
P(A = a|D) \propto \sum_{T} [P(A = a)P(T = t)] P(D|T = t, A = a) \tag{2}
\]

The attribution process tries to predict an author given an anonymous document, and that can be shown in Eq. 2 which follows from Eq. 1 after applying Bayes rule. The same argument about the topic also applies to the language model, but for simplicity, we only focus on the topic since POS tags have been shown to capture the stylistic variations in language grammar between authors.

Same-topic scenarios, assumes that the document generation depends only on the author’s writing choices, completely ignoring the topic, i.e., no \(T\) in the joint distribution, and so, \(P(A = a|D)\) is \(\propto P(A = a)P(D|A = a)\), but this is unrealistic and unintuitive. In contrast, cross-topic scenarios account for the topics’ effects but they assume that the topic is independent from the author. This is clear from the cross-topic setup where the topic values are fixed during training and testing. While this setup managed to reduce the effect of the topic on the attribution process, it does not help in identifying the causes of the errors that resulted from changing the topic between training and testing.

Instead, we propose a setting in which the topic is dependent on the author, as shown in Figure 2b. Our intuition about the effect of the author’s writing style on the topic is the following. If we consider that each topic has its unique distribution over words, which may be captured if we analyze a fairly large corpus of documents on that topic, then this distribution will be slightly different in one single document written by a certain author. This is because the limited document’s length will force an author to choose a subset of words that describe that specific topic. This topic distribution will also differ across documents written by different authors because words have synonyms and the same idea can be worded in multiple ways.

If we allow the topic to depend on the author, then the joint distribution changes from Eq. 1 to Eq. 3, and the conditional probability of an author given the anonymous document will change to Eq. 4.

\[
P(A, T, D) = P(A)P(T|A)P(D|A, T) \tag{3}
\]

\[
P(A = a|D) \propto \sum_{T} [P(A = a)P(T = t|A = a)] P(D|T = t, A = a) \tag{4}
\]

Because we allow the topic to depend on the author, we can create a scenario where a learning algorithm only sees samples on one topic for a specific author in the training set but a different topic in the test set, then we measure the error caused by this switch. Note that this proposed scenario will not be as easy as the same-topic, respects the restriction of the cross-topic scenario, and can help us understand the entanglement of the topic and the writing style.

### 3.2 The Proposed Setup

We propose a new task to measure the performance of authorship attribution techniques while being confused by the topics in the authors’ writing samples. The key component of this task is how we split the writing samples into training, validation and testing sets. To begin with, we define the notion of a cross-topic dataset as the set of writing samples written by \(N\) authors on \(T\) topics where the number of authors \(N \geq 4\), the number of topics \(T \geq 3\), and each author has, approximately, the same number of writing samples on each one of the \(T\) topics. First, we divide the authors into two equal-size groups: group 1 and group 2. Next to create the training set, we select two random topics and use writing samples on topic 1 for the authors in group 1 and writing samples on topic 2 for the authors in group 2. For the testing set, we flip the topics configuration that we used for the training set. We use writing samples on topic 2 (instead of 1) for the authors in group 1 and samples on topic 1 (instead of 2) for the authors in group 2. Finally, we use the remaining writing samples on the unused topics for the authors in both groups for the validation set. Figure 3 shows the setup for the proposed task as an example of having four authors and four topics.

After creating the training, validation, and testing sets we perform the typical model-based authorship attribution. First, the features are extracted from the writing samples. Second, a classification model is trained on the training samples, tuned on
Figure 3: Topic confusion task. We use two topics for training and switch them for testing. Two topics are used for hyperparameter tuning. The topic labels are not available for the classifier during training, and are only used to distribute the samples over the subsets and calculate the scores.

the validation set to pick the best hyperparameters, and tested on the testing set. Note that the classifier does not have access to any information about the setup, such as the groups configuration or the topic labels. Finally, instead of looking only at the classification accuracy, we take a closer look at the misclassified samples and propose the following simple measures to evaluate the performance.

**Correct**: The number of correctly classified samples. Dividing this number by the total number of predicted samples is the model’s unweighted classification accuracy.

**Same-group error**: The number of misclassified samples to authors within the same group as the true author.

**Cross-group error**: The number of misclassified samples to authors in the other group. By switching the topics between the training and the testing sets, we try to confuse the attribution model. Our hypothesis is as follows. Features that are invariant to the topic and only capture the authors’ writing styles should lead a model to classify the samples to their true authors and this can be measured by counting the number of correctly classified samples. Alternatively, features that capture the topic instead of the writing styles would lead a model to follow the topics and, as a result, classify the samples to authors in the opposite group which has the same topic that the model saw during training. We propose the cross-group error where we count the number of misclassified samples to authors in the opposite group, and not the same group as the true author. Finally, a model that cannot distinguish the authors’ writing styles—will misclassify samples to authors within the same group. To measure that, we use the same-group error where we count the number of misclassified samples to authors within the same group of the true author.

Compared to the standard cross-topic setting, this task can help understand how the topic affects certain features by showing whether the error is caused by the topic or the features themselves, while the cross-topic setting would give a more realistic performance compared to the same-topic, but without any insights on why we got such results.

4 Dataset

We present an extended, curated, and fairly balanced version of the Guardian dataset. One motivation is that as we try to understand the effect of the topic on the attribution process, we need to isolate any external factors that may affect the performance and make the results noisy. For example in the topic confusion task, we have to use topics that have writing samples from all the authors. Otherwise, the model could learn to favor one topic versus the other during training, and on test time will have author samples that it did not see during training. Based on that, it will be hard to tell whether these samples are going to be misclassified due to lack of training samples or due to a strong topic effect on the attribution process. Although datasets in real-life can be imbalanced, this issue can be addressed by either randomly excluding some writing samples to make the dataset imbalanced, or by using proper performance metrics for imbalanced datasets such as weighted accuracy, precision and recall or F-Score.
Another reason is that the articles in the commonly used version of the dataset contained some HTML artifacts and meta-data from the Guardian’s website, and had a number of its articles either on the wrong topic, or written by authors that are not in the dataset. Because of that, we retrieved the original articles, and added more articles to balance the number of writing samples per author on each topic. We maintained the same upper limit on the number of documents per author as the original dataset. The data collection procedure is described in Appendix A. The total number of collected articles are provided in Table 1.

| # of articles/author | # of articles/topic |
|----------------------|---------------------|
| M.K. 35              | Politics (P) 130    |
| H.Y. 37              | Society (S) 118*    |
| J.F. 38              | UK (U) 130          |
| M.R. and P.P. 39     | World (W) 130       |
| The remaining 8      |                     |

Table 1: The number of articles per topic and per author in our dataset. Descriptive statistics are provided in Appendix D. Table 3 (* Has less than 10 articles/author)

5 Experimental Setting

5.1 Quantifying Writing Style

Stylometric Features We adopted the stylometric features used in (Iqbal et al., 2008) which were originally based on the work of Abbasi and Chen (2006). In our work, we use the lexical features on both character- and word-level, with syntactic features. In our experiments we have 371 features and are described in Table 4 in Appendix B.

N-Grams Using n-grams is a common approach to represent documents in authorship attribution (Stamatatos, 2013; Sapkota et al., 2014, 2015). Note that as these features are used to represent the style, they can be referred to as stylometric features. However, researchers have distinguished this approach as all the features are extracted in the same way: tokenization first, then counting the tokens. For most text classification tasks the tokenization is done on either the word- or on the character-level in a text classification task.

Pretrained Language Models We used the BERT Devlin et al. (2019) and RoBERTa (Liu et al., 2019) pretrained sequence classification models that are provided by the HuggingFace (Wolf et al., 2019) library. These contextual language models consist of a transformer-based embeddings layer with 110 million and 125 million parameters, respectively, followed by a classification layer. Due to the huge size of these models and the small number of training samples we could not train the whole model, i.e., both embeddings and classification layers. Note that this behavior is expected where large models can overfit the training data until a certain point where the model is too huge and the data is too little for the training process to be effective (Nakkiran et al., 2019). To overcome this issue, we decided to freeze the embeddings and train only the classification layer. There are no hyperparameters for change the structure of these models as they have predetermined vocabulary size, and a maximum sequence length. We used the base, lower-case version of these models.

5.2 The Attribution Model

In this work, we adopt the instance-based approach (Stamatatos, 2009) where a writing style is extracted from every sample separately, and a model is trained on the writing samples to predict the authors of new, unseen, samples. We use Pedregosa et al. (2011)’s implementation of linear Support Vector Machines (SVM) as the classification algorithm, which is a common choice in authorship attribution (Stamatatos, 2017). The literature has thoroughly compared and contrasted the performance of different classification techniques such as Naïve Bayes, decision trees and SVM (Ding et al.,

2The supplementary material contains extensive details on the data collection process including which documents were excluded and the reason for exclusion, a list of the articles URLs, and the code to scrape and preprocess the articles to the format which we used.

3We used the POS tagger from (Manning et al., 2014).

proven to be an important indication of style (Sundararajan and Woodard, 2018).

Masking This preprocessing technique replaces every digit with the (#) symbol, and every character of the words that will be masked with the (*) symbol. Masked words are chosen based on their frequency in an external dataset, namely, the British News Corpus (BNC). After Masking, tokens are put back together to recreate the original document and paragraphs structure. Next, n-grams are used either on the word- or on the character-level in a text classification task.

4https://huggingface.co
We optimize the following hyperparameters. $k$ is the threshold for masking, $n_w$ is the word-level and POS $n$-grams, $n_{ch}$ is the char. level $n$-gram, and $f_l$ is the minimum frequency threshold in the whole dataset. In Appendix C, Tables 5 shows the range of values and 6 shows the average optimal parameters that are fine-tuned on the validation set.

5.4 Evaluation Procedure

For each set of features, we use the setup explained in Section 3 and report the average score for a 100 runs. In each run, we randomize the order of the topics, the authors used, and the distribution of the authors on the groups. This is considered as one single experiment. To reduce the variance in the results, each single experiment was repeated ten times,$^5$ and we reported the average balanced accuracy score.

6 Results and Discussion

In this section, we evaluate the performance of different combination of the features described in Sec. 5.1 in the proposed topic confusion task, and in the cross-topic setting. Additionally, we use this task to support the intuition of early researchers on authorship attribution who used function words and other stylometric features to represent the writing style. Table 2 shows the results when using these features alone, combined with stylometric features, or with both stylometric features and POS $n$-grams.

The first important observation from Table 2 is how poorly pretrained language models perform on this task. Indeed, pretrained neural language models are not suited for authorship attribution because their goal is to learn word representations that capture the similarity between words that appear in a similar context, something that one-hot encoding does not capture. In authorship attribution, however, words must both appear in the same context and be used by that author to have similar representations. To illustrate this, consider the words ‘color’ and ‘colour’, which are the same word but with different spelling based on whether American or British English is being used. Ideally, these two words would have a very similar embeddings if not identical ones. In authorship, however, the distinction between the two is very important because it indicates the author’s identity or at least the language system that they use. Authorship attribution techniques try to highlight these differences and use them to identify the most plausible author of an anonymous document.

One reason why these language models performed worse than word-level $n$-grams is the difference in the vocabulary set. The vocabulary set in these pretrained models is decided based on an external, much larger dataset which usually excludes words with a frequency less than 10 or 20 in that dataset. For $n$-grams, however, the vocabulary set is chosen based on the investigated dataset.

Another main observation from Table 2 is that combining either stylometric features alone or both stylometric features and POS $n$-grams with all the other features has decreased the cross-group error significantly which resembles less confusion over the topic and better capturing of the authors writing styles. On the other hand, the same-group error was reduced by merely one sample at max in most of cases except when stylometric features are combined with POS $n$-grams.

Similarly, we can see the benefit of masking infrequent words (using the masking approach) then using tokenization is evident in the cross-group error between character $n$-grams and masking on the character-level as well as the word $n$-grams and masking on the word-level. As we can see in Table 2, same-group error remained fixed while cross-group error decreased by 11 and 17 samples for the character- and the word-level, respectively.

We can also see that stylometric features alone has the worst same-group error which means that they cannot fully capture the authors writing styles despite being less affected by the topic. This agrees with the literature on same-topic authorship attribution where several studies have shown that stylometric features can be easily outperformed by character, word or POS $n$-grams. In contrast, word- and character-level $n$-grams have the worst cross-group error scores which is reasonable because, naturally, these features are used for topic classification.

When comparing character- and word-level $n$-grams, we see that they have very similar same-group errors while cross-group error is much higher for word $n$-grams. In the classical cross-topic authorship scenario, literature has shown that character $n$-grams outperform word $n$-grams while still

---

$^5$We trained BERT and RoBERTa only once.
Table 2: Results on the topic confusion task and the cross-topic scenario. The first row of each group of rows corresponds to an existing method. The last row is the performance at random. **Boldface** indicates the best result per column. (↑ Higher is better. ↓ Lower is better. S: number of samples. %: Percentage. *State of the art.)

| Features                  | ↑ Accuracy (%) | ↑ Correct (S) | ↓ Same-group Error (S) | ↓ Cross-group Error (S) | ↑ Accuracy (%) |
|---------------------------|----------------|---------------|------------------------|------------------------|----------------|
| **Topic Confusion**       |                |               |                        |                        |                |
| Stylo                     | 62.7           | 73.8          | 18.3                   | 24.8                   | 61.2           |
| POS n-grams               | 71.4           | 84.2          | 13.4                   | 19.4                   | 71.0           |
| + Stylo                   | 79.2           | 93.1          | 9.8                    | 14.1                   | 79.2           |
| Char n-grams              | 69.6           | 82.0          | 7.9                    | 27.1                   | 77.3           |
| + Stylo                   | 72.6           | 85.4          | 7.6                    | 24.0                   | -              |
| + Stylo & POS             | 76.4           | 89.8          | 7.0                    | 20.1                   | 82.8           |
| Word n-grams              | 62.2           | 73.1          | 9.3                    | 34.6                   | 77.7           |
| + Stylo                   | 74.6           | 87.7          | 8.6                    | 20.6                   | -              |
| + Stylo & POS             | 80.0           | 93.9          | 8.3                    | 14.7                   | **83.3**       |
| Masking (Ch.)*            | 79.1           | 92.9          | 7.9                    | 16.1                   | 80.9           |
| + Stylo & POS             | 82.8           | 97.2          | 7.5                    | 12.2                   | 83.2           |
| **Cross-topic**           |                |               |                        |                        |                |
| Masking (W.)              | 76.3           | 89.8          | 9.3                    | 17.9                   | 77.9           |
| + Stylo & POS             | **83.0**       | **97.5**      | 7.8                    | **11.7**               | 82.8           |
| BERT-base                 | 32.9           | 38.7          | 23.3                   | 55.0                   | 37.5           |
| RoBERTa-base              | 39.5           | 46.6          | 15.3                   | 55.1                   | 51.1           |
| “random chance”           | 8.3            | 9.4           | 48.7                   | 58.5                   | 7.7            |

capturing the topic. This is in line with our results as character n-grams have lower cross-group error, which makes them less influenced by the topic in the attribution task.

For the masking approach, we see an increase in the performance similar to what is observed with all the other features. It is worth noting that masking on the character-level performed better than the masking on the word-level without any additional features. In contrast, when stylometric features and POS n-grams are added to both masking approaches, the word-level one performed better.

Finally, as shown in the right-most column in Table 2, the previous state of the art on the cross-topic scenario is the character-level n-grams with masking. This technique is outperformed by combining stylometric features and POS n-grams with any other n-gram-based features regardless if masking was used for preprocessing or not. However, the difference in the performance was statistically significant only when masking (Ch.) is compared to the combination of word-level n-grams, POS n-grams and stylometric features ($P = 0.04$).

We briefly discussed the cross-topic results here, but we provide more details in Appendix D which includes the experimental setup (Appendix D Sec.1), detailed results and analysis supported with statistical significance tests (Appendix D Sec.2), and an ablation study on the cross-topic scenario (Appendix D Sec.3). Additionally, we evaluate the best performing features under the cross-topic scenario on the same-topic setup in Appendix D Sec.4, with detailed result on each dataset configuration in Appendix E, in Tables 10, 11 and 12.

7 Conclusion

In this work, we proposed the topic confusion task which is useful for testing the capacity of features in capturing the writing style while being influenced by the topic in authorship attribution. Additionally, it could help in understanding the cause of the errors in authorship attribution. We verified the outcomes of this task on the cross-topic authorship attribution scenario, and showed that stylometric features and POS tags can better capture the writing style compared to the commonly used n-grams. We achieved a new state of the art of 83.3% on the cross-topic scenario by resurrecting stylometric features and combining them with POS tags and word-level n-grams which is around 3% over the previous, state-of-the-art, masking-based, character-level approach.
References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Ahmed Abbasi and Hsinchun Chen. 2006. Visualizing authorship for identification. In *International Conference on Intelligence and Security Informatics*, pages 60–71. Springer.

Malik H. Altakrori, Farkhund Iqbal, Benjamin C. M. Fung, Steven H. H. Ding, and Abdallah Tubaishat. 2018. Arabic authorship attribution: An extensive study on twitter posts. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(1):5.1–5.51.

Douglas Bagnall. 2015. Author identification using multi-headed recurrent neural networks. In *Working Notes of Conference and Labs of the Evaluation Forum (CLEF)*.

Georgios Barlas and Efstatios Stamatakos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, pages 255–266. Cham. Springer International Publishing.

Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-language authorship attribution. In *Proc. of the Ninth Int’l Conf. on Language Resources and Evaluation (LREC)*, pages 2015–2020, Reykjavik, Iceland. European Language Resources Association (ELRA).

José Eleandro Custódio and Ivandré Paraboni. 2019. An ensemble approach to cross-domain authorship attribution. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 201–212. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Steven H. H. Ding, Benjamin C. M. Fung, and Mourad Debbabi. 2015. A visualizable evidence-driven approach for authorship attribution. *ACM Trans. Inf. Syst. Secur.*, 17(3):12:1–12:30.

Steven H. H. Ding, Benjamin C. M. Fung, Farkhund Iqbal, and William K. Cheung. 2019. Learning stylistic representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

Georgia Frantzeskou, Efstatios Stamatakos, Stefanos Gritzalis, Carole E Chaski, and Blake Stephen Howald. 2007. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, 6(1):1–18.

Zhenhao Ge, Yufang Sun, and Mark JT Smith. 2016. Authorship attribution using a neural network language model. In *AAAI*, pages 4212–4213.

Jade Goldstein, Ransom Winder, and Roberta Sabin. 2009. Person identification from text and speech genre samples. In *Proc. of the 12th Conference of the European Chapter of the ACL (EACL)*, pages 336–344.

Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. 2018. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756.

Ana Granados, David Camacho, and Francisco Borja Rodríguez. 2012. Is the contextual information relevant in text clustering by compression? *Expert Systems with Applications*, 39(10):8537–8546.

Julian Hitschler, Esther van den Berg, and Ines Rehbein. 2017. Authorship attribution with convolutional neural networks and POS-eling. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58, Copenhagen, Denmark. Association for Computational Linguistics.

John Houvardas and Efstatios Stamatakos. 2006. N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 77–86. Springer.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Farkhund Iqbal, Hamad Binsalleeh, Benjamin C. M. Fung, and Mourad Debbabi. 2013. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112.

Farkhund Iqbal, Rachid Hadjidj, Benjamin C. M. Fung, and Mourad Debbabi. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5:S42–S51.

Patrick Juola. 2008. *Authorship Attribution*, volume 1. Now Publishers, Inc.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proc. of the Conference Pacific Association for Computational Linguistics, PACLING*, volume 3, pages 255–264.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Robert Layton, Stephen McCombie, and Paul Watters. 2012. Authorship attribution of irc messages using inverse author frequency. In *Proc. of the 3rd Cybercrime and Trustworthy Computing Workshop (CTC)*, pages 7–13.
Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*, pages arXiv–1907.

Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing*, 26(1):35–55.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237–249.

George K. Mikros and Eleni K. Argiri. 2007. Investigating topic influence in authorship attribution. In *Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*. CEUR-WS.org.

Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. 2019. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*.

James O’Shea. 2013. Alphabetical order 277 word new function word list. https://semanticsimilarity.files.wordpress.com/2013/08/jim-oshea-fwlist-277.pdf. [Retrieved Oct. 2019].

Rebekah Overdorf and Rachel Greenstadt. 2016. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proc. on Privacy Enhancing Technologies*, 2016(3):155–171.

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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, David Pinto, and Liliana Chanona-Hernández. 2017. Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3):627–639.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilou, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatakis. 2016. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686*.

Miguel A Sanchez-Perez, Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. 2017. Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In *Proc. of the Int’l Conf. of the Cross-Language Evaluation Forum for European Languages*, pages 145–151. Springer.

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.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In *Proc. of the 25th International Conference on Computational Linguistics (COLING)*, pages 1228–1237, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yunita Sari, Andreas Vilchos, and Mark Stevenson. 2017. Continuous n-gram representations for authorship attribution. *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, page 267.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proc. of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 1880–1891.
Prasha Shrestha, Sebastian Sierra, Fabio A González, Paolo Rosso, Manuel Montes-y Gómez, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), page 669.

Grigori Sidorov, Francisco Velasquez, Efstatios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. Expert Systems with Applications, 41(3):853–860.

Rui Sousa Silva, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio Oliveira, and Belinda Maia. 2011. "twazn me!!!;(' automatic authorship analysis of micro-blogging messages. In Natural Language Processing and Information Systems, pages 161–168. Springer.

Efstatios Stamatatos. 2009. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3):538–556.

Efstatios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy, 21:421–439.

Efstatios Stamatatos. 2017. Authorship attribution using text distortion. In Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics, volume 1, pages 1138–1149.

Efstatios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. Journal of the Association for Information Science and Technology, 69(3):461–473.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents “style” in authorship attribution? In Proceedings of the 27th International Conference on Computational Linguistics, pages 2814–2822.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface’s transformers: State-of-the-art natural language processing. ArXiv, abs/1910.03771.

Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. 2018. Syntax encoding with application in authorship attribution. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2742–2753.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Advances in neural information processing systems, pages 649–657.
A Data Collection

First, we curated the existing dataset by retrieving the 381 original documents from the Guardian’s website. Next, we inspected the authors’ names and the topics associated with each article. We excluded the articles that had the wrong topic (e.g. labelled as “Politics” in the dataset while having a “Society” tag on the website), or the ones that appeared under more than one of the previous topics, or were co-authored by multiple authors.

Next, we used the Guardian’s API\textsuperscript{1} to get all the articles written by each author, filtered them based on the topic, and collected the URLs of these articles and new articles aiming for 10 documents per author per topic. This resulted in a total of 40 documents per author. Note that while some authors have been writing in the Guardian for more than 20 years, they would mostly focus on one topic while occasionally writing on the other four. As a result, we still could not get 10 articles per author on the Society topic. The supplementary material contains full instructions, and the necessary script to get the data and preprocess it. We provide some descriptive statistics for the collected dataset in Table 3.

| Total number of: | Number of articles per topic |
|------------------|-----------------------------|
| Topics           | 4                           |
| Authors          | 13                          |
| Articles         | 508                         |
| Words            | 3,125,347                   |
| Politics (P)     | 130                         |
| Society (S)      | 118*                        |
| UK (U)           | 130                         |
| World (W)        | 130                         |

| Average number of: | Number of articles per author |
|--------------------|-------------------------------|
| Articles / Author  | $39.1 (SD = 1.5)$             |
| Articles / Topic   | $127 (SD = 5.2)$              |
| Words / Author     | $41K (SD = 6.9K)$             |
| Words / Topic      | $781K (SD = 13.0K)$           |
| Words / Document   | $1050.2$                      |
| M.K.               | 35                           |
| H.Y.               | 37                           |
| J.F                | 38                           |
| M.R. and P.P.      | 39                           |
| The remaining 8    | 40                           |

Table 3: Descriptive statistics for the extended Guardian dataset (* Has less than 10 articles per author).

B Stylometric Features

| Lexical Features - Character-Level | Lexical Features - Word-Level |
|-----------------------------------|------------------------------|
| 1. Characters count (N)           | 1. Tokens count (T)          |
| 2. Ratio of digits to N           | 2. Average sentence length (in characters) |
| 3. Ratio of letters to N          | 3. Average word length (in characters) |
| 4. Ratio of uppercase letters to N| 4. Ratio of alphabets to N   |
| 5. Ratio of tabs to N             | 5. Ratio of short words to T (a short word has a length of 3 characters or less) |
| 6. Frequency of each alphabet (A-Z), ignoring case (26 features) | 6. Ratio of words length to T. Example: 20% of the words are 7 characters long. (20 features) |
| 7. Frequency of special characters: $<=>%{}[\_\@\#\^\*\+=\$\&\_\()' (24 features). | 7. Ratio of word types (the vocabulary set) to T |

| Syntactic Features | |
|--------------------|---|
| 1. Frequency of Punctuation: , . ? ! : ; ’ ” (8 features) | |
| 2. Frequency of each function words (O’Shea, 2013) (277 features) | |

Table 4: List of stylometric features.

\textsuperscript{1}https://open-platform.theguardian.com
C  Optimal Hyperparameters

| Hyperparameter | Range                      |
|----------------|----------------------------|
| \(k\)          | 100, 200, 300, 400, ..., 500, 1000, 2000, 3000, 4000, 5000 |
| \(f_t\)        | 5, 10, 15, 20, 25, 30, 35, 40, 45, 50               |
| \(n_{ch}\)     | 3, 4, 5, 6, 7, 8                                |
| \(n_w\)        | 1, 2, 3                                      |

Table 5: Hyperparameters for masking and \(n\)-gram based feature representations. \(k\) is the threshold for masking, \(n_w\) is the word-level and POS \(n\)-grams, \(n_{ch}\) is the character-level \(n\)-gram, and \(f_t\) is the minimum frequency threshold in the whole dataset.

For pretrained BERT and RoBERTa, we used the pretrained sequence classification models. These pretrained models do not have hyperparameters for the model structure, but only have pretrained configurations. We used the base uncased models, where base refers to the models’ size (not large, and not distilled) and trained on all-lower-case text. For the training procedure, we used the following: AdamOptimizer, \(lr=0.1\), Epochs=500, EarlyStopping\((\text{min\_delta}=1e-3, \text{patience}=100)\). Despite the large Epoch value, most models would stop after less than 150 epochs.

| Method                  | \(k\) | \(n\) | \(f_t\) | Feat.          |
|-------------------------|-------|-------|---------|---------------|
| Masking (W.)            | 1,616.7 | 1.9   | 7.9     | 3,265.8       |
| Masking (Ch.)           | 1,691.7 | 5.5   | 18.8    | 6,416.3       |
| Stylometric + POS       | -     | 1.3   | 31.3    | 484.2         |
| Stylometric + POS + \(n\)-grams (W.) | -     | 2.0   | 12.5    | 2,481.0       |
| Stylometric + POS + \(n\)-grams (Ch.) | -     | 3.8   | 38.3    | 5,355.6       |

Table 6: The average optimal parameters for each feature representation, with the resulting number of features under these settings \((k\): masking threshold, \(n\): number of tokens in \(n\)-grams, \(f_t\): minimum frequency threshold in the dataset, W.: word-level, Ch.: character-level).

D  Additional Experiments

1  Data Splitting and Preprocessing

The Cross-Topic Scenario. In all our experiments, we split the dataset into training, validation and test sets. For the cross-topic experiments we followed the same setup in (Stamatatos, 2017). We used one topic for training, another topic for validation and hyperparameter tuning, and the remaining two topics for testing. The number of articles was 127 articles when training on Society and 130 articles otherwise. This setup resulted in 12 different topics permutations. We reported the average overall accuracy on all the 12 configurations.

The Same-Topic Scenario. We combined the 508 articles from all the topics, then split them as follows: 26\% for training, 26\% for validation, and the remaining 58\% for testing. This corresponds to 132 articles for training, 132 articles for validation, and 244 articles for testing. This ensures that the difference in performance between the same-topic and the cross-topic scenarios is not caused by the difference in the number of samples that are used for training/testing. We repeated this process 12 times and reported the average overall accuracy.

2  Cross-Topic Authorship Attribution

As shown in Table 7, by combining the stylometric features and POS tags with \(n\)-gram features we achieve the highest accuracy of 83.3\%. This is in line with our findings in the topic confusion task in Sec. 3. The difference between using all the features \((\text{mean} = 83.26, \text{SD} = 2.63)\) and the character-based masking approach \((\text{mean} = 80.89, \text{SD} = 2.59)\) is statistically significant \((P = 0.04)^{**}\).

**We used a t-Test: Two-Sample Assuming Unequal Variances at the \(\alpha = 0.5\) level.
We experimented with the previously proposed features for the same-case scenario. Table 9 shows why researchers considered them a good candidate for future representations. However, we show that combining them, as opposed to picking one versus the other can achieve higher accuracy than using each one of the separately.

| Features                  | Accuracy |
|---------------------------|----------|
| Stylo. + POS              | 79.2 ± (2.7) |
| Stylo. + POS + n-grams (W.) | **83.3 ± (2.6)** |
| Stylo. + POS + n-grams (Ch.) | 82.8 ± (2.7) |
| Masking (W.)              | 77.9 ± (4.0) |
| Masking (Ch.)             | 80.9 ± (2.6) |
| Masking (W.) + Stylo. + POS | 82.8 ± (3.3) |
| Masking (Ch.) + Stylo. + POS | 83.2 ± (3.3) |

Table 7: Average cross-topic classification accuracy (%) on the extended Guardian dataset (W.: word-level, Ch.: character-level).

| Features                  | Accuracy |
|---------------------------|----------|
| Stylo.                    | 61.2 ± (3.1) |
| POS                       | 71.0 ± (3.2) |
| W. n-grams                | 77.7 ± (2.7) |
| Ch. n-grams               | 77.3 ± (2.8) |
| Stylo. + POS              | 79.2 ± (2.7) |
| Stylo. + POS + n-gram (W.) | **83.2 ± (2.6)** |

Table 8: Ablation study: classification accuracy (%) on cross-topic scenario. (Stylo.: Stylometric, W.: word-level)

It is also worth noting that by using only stylometric features with POS n-grams we can achieve similar results to the masking approach with character-level tokenization. The difference of 1.7% in favor of the masking approach is statistically insignificant ($P = 0.15$) with a ($mean = 80.89, SD = 2.59$) for masking versus a ($mean = 79.22, SD = 2.70$) when using stylometric features with POS n-grams.

Furthermore, Table 7 shows a 3% increase in the accuracy for the masking approach when using character-level tokenization. This outcome is in line with the findings in (Stamatatos, 2017). The difference between word-level n-grams ($mean = 77.90, SD = 4.03$) and character-level ($mean = 80.89, SD = 2.59$) is statistically insignificant ($P = 0.05$). Similarly, the difference between combining the stylometric features and POS-grams with word-level n-grams ($mean = 83.26, SD = 2.63$) versus with character-level n-grams ($mean = 82.83, SD = 2.7$) is statistically insignificant ($P = 0.71$).

Finally, the difference between the state-of-the-art approach which is masking on the character-level from one side, versus stylometric features and POS tags combined with either character-level n-grams ($mean = 80.89, SD = 2.59$), masking on the word-level ($mean = 82.80, SD = 3.34$) or masking on the character-level ($mean = 83.17, SD = 3.33$) is statistically insignificant ($P = 0.10, P = 0.98$, and $P = 0.80$, respectively.). The only statistically significant difference ($P = 0.04$) was with stylometric features and POS tags combined with word-level n-grams ($mean = 83.26, SD = 2.63$).

3 Ablation Study on the Cross-Topic Scenario

We conclude our experiments with an ablation study to see the contribution of each set of features to the overall accuracy. Similar to the experiments above, we perform a grid search over all the hyperparameters $f_l$ and $n$. Table 8 shows the results.

As shown in this experiment, each feature set on its own does not achieve the same performance as with combining all of them. We also confirm the previous results where, even in the cross-topic scenario, n-grams outperformed stylometric features by a large margin.

We evaluated the significance of the difference between the top three accuracy groups. The results show that the difference between Set (3) ($mean = 77.7, SD = 2.69$) and Set (4) ($mean = 79.3, SD = 2.7$) is statistically insignificant ($P = 0.21$) while it is significant ($P < 0.01$) between Set (4) and Set (5) ($mean = 83.3, SD = 2.6$).

4 Same-Topic Authorship Attribution

We experimented with the previously proposed features for the same-case scenario. Table 9 shows that, in line with the literature, n-gram based features outperform stylometric features which justifies why researchers considered them a good candidate for future representations. However, we show that combining them, as opposed to picking one versus the other can achieve higher accuracy than using each one of the separately.
| Method                                      | Accuracy       |
|--------------------------------------------|----------------|
| Stylo.                                     | 64.2 ± (2.3)   |
| n-grams (W.)                               | 79.1 ± (2.3)   |
| Masking (W.)                               | 81.7 ± (3.1)   |
| Stylo. + POS + n-grams (W.)                | 87.5 ± (2.6)   |

Table 9: Average same-topic classification accuracy (%) on the extended Guardian dataset. (Stylo.: stylometric, W.: word-level, Ch.: character-level)

E Case-by-Case Attribution Accuracy

| Topics       | Accuracy (%)                     |
|--------------|----------------------------------|
|              | Masking                         | Stylometric + POS | Sty. + POS + n-grams |
|              | Word lvl. | Char. lvl. | Word lvl. | Char. lvl. | Word lvl. | Char. lvl. |
| P S U & W    | 76.1 ± 1.1 | 80.9 ± 0.9 | 80.0 ± 1.2 | 85.4 ± 0.9 | 80.0 ± 1.2 |
| P U S & W    | 81.8 ± 1.1 | 86.5 ± 0.7 | 84.2 ± 1.2 | 87.4 ± 1.2 | 87.2 ± 0.9 |
| P W S & U    | 79.6 ± 0.7 | 81.6 ± 1.6 | 79.6 ± 1.1 | 87.6 ± 0.7 | 82.5 ± 1.0 |
| S P U & W    | 70.6 ± 1.4 | 76.3 ± 0.9 | 76.0 ± 1.9 | 78.8 ± 1.4 | 77.9 ± 1.3 |
| S U P & W    | 76.6 ± 1.7 | 79.0 ± 1.9 | 77.0 ± 1.4 | 82.4 ± 1.2 | 82.8 ± 0.6 |
| S W P & U    | 71.2 ± 1.3 | 81.2 ± 1.5 | 75.9 ± 1.8 | 81.3 ± 1.6 | 83.0 ± 0.9 |
| U P S & W    | 80.4 ± 1.5 | 79.3 ± 1.0 | 79.3 ± 0.9 | 82.1 ± 1.7 | 82.1 ± 1.3 |
| U S P & W    | 81.2 ± 1.2 | 79.0 ± 1.6 | 79.7 ± 1.1 | 83.7 ± 1.0 | 83.9 ± 1.5 |
| U W P & S    | 83.7 ± 1.8 | 80.2 ± 1.2 | 83.7 ± 1.3 | 85.4 ± 0.8 | 84.6 ± 0.9 |
| W P S & U    | 74.9 ± 1.9 | 81.4 ± 0.8 | 75.6 ± 1.5 | 79.9 ± 1.2 | 79.1 ± 0.8 |
| W S P & U    | 77.0 ± 1.9 | 80.4 ± 1.0 | 78.9 ± 1.3 | 82.1 ± 1.1 | 84.1 ± 0.9 |
| W U P & S    | 81.6 ± 1.4 | 84.9 ± 1.0 | 80.7 ± 1.2 | 83.1 ± 1.4 | 86.7 ± 0.7 |
| Average      | 77.9       | 80.9       | 79.2       | 83.3       | 82.8       |

Table 10: Cross-topic classification accuracy (%) ± SD on the extended Guardian dataset.
### Table 11: Ablation study: classification accuracy (%) on cross-topic scenario.

| Topics       | Accuracy (%) | n-grams (W.) | Stylo. + POS + n-gram (W.) |
|--------------|--------------|--------------|----------------------------|
| Train. Valid. Test. | Stylo. | POS | 64.1 ± 2.9 | 72.8 ± 1.2 | 74.7 ± 1.2 | 80.0 ± 1.2 |
| P S U & W    | 66.2 ± 2.7  | 76.9 ± 1.1  | 82.9 ± 0.7  | 84.2 ± 1.2 |
| P W S & U    | 65.7 ± 2.9  | 69.5 ± 1.0  | 76.4 ± 1.5  | 79.6 ± 1.1 |
| S P U & W    | 57.4 ± 2.7  | 66.9 ± 1.2  | 72.7 ± 0.9  | 76.0 ± 1.9 |
| S U P & W    | 57.4 ± 3.0  | 68.7 ± 1.2  | 79.0 ± 1.4  | 77.0 ± 1.4 |
| S W P & U    | 57.7 ± 2.1  | 65.1 ± 1.5  | 76.7 ± 0.6  | 75.9 ± 1.8 |
| U P S & W    | 60.4 ± 2.0  | 72.9 ± 1.6  | 77.9 ± 2.0  | 79.3 ± 0.9 |
| U S P & W    | 60.0 ± 2.5  | 70.6 ± 0.8  | 80.7 ± 1.6  | 79.7 ± 1.1 |
| U W P & S    | 63.3 ± 2.7  | 74.7 ± 1.6  | 80.5 ± 0.8  | 83.7 ± 1.3 |
| W P S & U    | 60.3 ± 3.8  | 69.2 ± 1.7  | 75.8 ± 1.9  | 75.6 ± 1.5 |
| W S P & U    | 58.3 ± 3.4  | 72.7 ± 1.3  | 76.9 ± 0.6  | 78.9 ± 1.3 |
| W U P & S    | 63.9 ± 2.9  | 71.8 ± 1.3  | 78.5 ± 0.8  | 80.7 ± 1.2 |
| Average      | 61.2         | 70.1         | 77.7         | 79.2         | 83.3         |

### Table 12: Classification accuracy (%) on same-topic scenario.

| Run | Accuracy (%) |
|-----|---------------|
|     | n-grams (W.) | Masking (W.) | Stylo. | Stylo. + POS + n-gram (W.) |
| 1.  | 81.6 ± 1.6    | 80.8 ± 0.8   | 60.9 ± 2.1 | 85.0 ± 0.6 |
| 2.  | 81.0 ± 1.8    | 81.5 ± 0.9   | 60.7 ± 2.9 | 88.8 ± 1.1 |
| 3.  | 79.8 ± 1.5    | 78.5 ± 1.6   | 65.9 ± 2.9 | 88.0 ± 0.9 |
| 4.  | 79.4 ± 1.5    | 82.1 ± 1.1   | 63.6 ± 1.5 | 88.2 ± 1.8 |
| 5.  | 77.1 ± 0.9    | 79.7 ± 1.7   | 66.7 ± 2.5 | 89.2 ± 1.7 |
| 6.  | 80.1 ± 1.5    | 87.2 ± 1.3   | 68.6 ± 1.6 | 91.3 ± 0.8 |
| 7.  | 77.5 ± 0.8    | 78.9 ± 1.6   | 64.4 ± 3.7 | 86.2 ± 1.4 |
| 8.  | 79.2 ± 1.6    | 87.8 ± 1.8   | 62.8 ± 3.7 | 92.3 ± 1.1 |
| 9.  | 79.4 ± 0.8    | 78.3 ± 1.4   | 64.1 ± 2.8 | 84.8 ± 2.2 |
| 10. | 74.0 ± 1.6    | 80.6 ± 1.8   | 62.5 ± 2.2 | 82.8 ± 1.2 |
| 11. | 83.2 ± 1.1    | 84.7 ± 1.3   | 66.8 ± 2.9 | 87.0 ± 0.9 |
| 12. | 77.4 ± 1.6    | 80.6 ± 1.1   | 62.9 ± 3.9 | 85.9 ± 1.5 |
| Average | 79.1 | 81.7 | 64.2 | 87.5 |