FOI Cross-Domain Authorship Attribution for Criminal Investigations

Notebook for PAN at CLEF 2019

Fredrik Johansson and Tim Isbister

Swedish Defence Research Agency (FOI)
Stockholm, Sweden
{fredrik.johansson, tim.isbister}@foi.se

Abstract Authorship attribution techniques have existed for a long time, but they are seldom evaluated in conditions similar to the real-world scenarios in which they have to work if they should be useful tools in criminal investigations involving digital communication. We have used a SVM classifier as a base, onto which we have added two sets of hand-crafted stylometric features and evaluated it using data from the PAN-CLEF 2019 cross-domain authorship attribution task. Results outperform the baseline systems to which our classifiers have been compared.

1 Introduction

Despite that the uniqueness of individuals' fingerprints has been widely known for several decades, collecting and matching fingerprints from crime scenes to the fingerprints of suspects or large databases is still a useful tool in many criminal investigations. However, an increasing amount of crimes are carried out in the digital environment rather than in the physical world. People are sending threatening e-mails to politicians. Drug sellers advertise for cheap LSD on illegal darknet marketplaces. Terrorists post violent extremism propaganda on social media. The list goes on and on. In common for many crimes in the digital arena is that they in many cases involve digital written communication of various types. Therefore, it is unsurprisingly that many researchers in recent years have considered the idea to "fingerprint" online users by extracting and, in various ways, compare stylometric features such as distribution of function words, parts of speech, and word lengths from their written texts. One well-known and promising example of this type of methods is the Writeprint technique, developed by researchers at the University of Arizona [1]. In their experiments, as well as in many other authorship attribution studies, it is often assumed that criminal investigators have access to texts written by a (usually quite small) set of candidate authors, and that the anonymous texts with unknown authorship have been written by one of these candidate authors. Different studies have shown promising results for many types of digital communication, involving everything from e-mails [15] and forum posts [11] to tweets [9] and

blog posts [10]. One fundamental issue with these studies is that they tend to almost always be conducted on English texts. Far from all offenders in the digital environment communicate in English, so there is a need for more studies on how this kind of stylometric techniques work for other languages. Another issue is that criminal investigators in practice far from always are able to access training data from the same type of written communication as the anonymous texts that are supposed to be matched against the training data. However, it is quite seldom that studies evaluate how well authorship attribution algorithms perform when training on data from other medium types than they are tested on. Moreover, the available texts are seldom controlled for topic, leading to possibilities that authors may sometimes be distinguished based on topic rather than style. Given these issues it is still not obvious whether stylometric techniques are practically useful for law enforcement in real-world criminal investigations or not. For this reason, the PAN 2019 cross-domain authorship attribution task [7] is highly interesting. First of all it is cross-domain, meaning that texts of known and unknown authorship come from different domains. It is also the case that the challenge is an example of open-set attribution, meaning that the true author is not necessarily included in the list of candidate authors. Finally, it is also multilingual in the sense that it in addition to English also covers French, Italian, and Spanish texts.

The rest of this paper is structured as follows. First, we give an overview of previous work on authorship attribution in Section 2. In Section 3, we briefly describe the data available for the PAN 2019 authorship attribution task and present different ideas for how the challenge at hand can be tackled using different overall design choices. In Section 4, we describe the features we have implemented and used in our submitted systems. This is followed by Section 5, in which we give a detailed system description of how the various features have combined, scaled, and fed into two different classifiers. The obtained results are presented in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

A systematic review of the research field is outside the scope of this paper, but it is important to note that authorship attribution overall is a quite well-studied problem. As mentioned by Juola in his review of the history of the authorship attribution research field forward to around year 2006, the idea of telling something about a person from the language he or she uses goes back at least until the days of the Old Testament [6]. As clearly stated by Juola, the underlying theory behind most authorship attribution approaches is that some kind of fingerprint of an author can be extracted as a summary statistic from a given text, and that different authors will vary, noticeably and consistently, along this summary statistic. In his survey, Juola identifies the open class version of the problem (in which the true author is not necessarily in the set of identified candidate authors) to be much harder than the closed class version, and stresses the importance and difficulty of developing accurate cross-genre techniques. [6]. Another good overview of the authorship attribution field is given in [13]. Among other things, Stamatatos mentions the importance of the open class version of the problem as well as the importance of attribution methods to be robust even given a limited amount of rather

short texts [13]. He also highlights the need of finding specific stylometric features which capture only authorial information, rather than genre or topic. Much research has been undertaken since these overviews were conducted. One interesting example is the Writeprint technique introduced by Abbasi and Chen [1]. Another study worth mentioning is that by Narayanan et al. [10], in which the authors show that authorship attribution can be conducted also on very large scale with good accuracy.

3 Overall Design Choices

A detailed description of the PAN 2019 cross-domain authorship attribution dataset can be found in [7], but we here provide a short explanation of the most important aspects as they have an impact on how we have approached the problem in our implementations.

Participants in the challenge have got access to a development corpus with highly similar characteristics as an unseen evaluation corpus (to which we have not got access, but on which the developed system has been evaluated by submitting it and running it on TIRA [12]). Both the development and evaluation corpora consist of a set of cross-domain authorship attribution problems in each of the following languages: English, French, Italian, and Spanish. A fundamental restriction to note is that the sets of candidate authors of the development and the evaluation corpora are not overlapping, so that it is not possible to pre-train a model for classifying the authors of interest on the development corpus and simply apply it to the evaluation corpus. This constraint made us realize that we either have to:

1. Develop a solution which dynamically builds a (supervised) model on-the-fly as it encounter new problem instances, or
2. Develop a more general solution which learns whether two texts are likely to have been written by the same author, and apply this to all authors in each new problem instance.

Since we have previous experience of developing and applying the second type of techniques to the related problem of author verification [2], [5], we started by applying a similarity-based supervised classifier [2] intended to classify pairs of texts as to whether they have being written by the same author or not. When dealing with such a small number of text documents for each author as were present in the development corpus, our initial impression was that a more general solution to the problem might perform better. However, our existing classifier had only been developed for English and Swedish and has been trained on forum posts, which is quite different from the data used in this challenge. Since initial experimental results were not very encouraging, we decided not to spend the time on collecting more representative data to train this kind of binary classifiers on for all the languages of interest. Instead, our intuition on this point suggested that training a more tailored supervised classifier dynamically for each new problem instance would be a more viable approach.

We got many initial ideas for what type of supervised classifiers to build, many of them involving various kinds of deep learning architectures. Everything from making use of a pre-trained language model per language which could be fine-tuned to each individual candidate author, to building Siamese networks were up on the drawing board.

However, after making a few tests on the TIRA [12] platform we realized that the virtual machines and their lack of GPUs would make this kind of dynamic model building per problem instance take forever using such advanced architectures. For this reason, we have in the end decided to go with a much more "traditional" authorship attribution approach. This approach has many similarities with the baseline-SVM implementation provided by the PAN organizers. However, it has been complemented with many hand-crafted stylometric features.

4 Features

This section presents the full list of hand-crafted features that have been implemented in any of our submitted solutions. Overall, stylometric features are intended to reflect stylistic characteristics of the writing of individual authors. The idea is that they should be as independent of topic as possible, and instead capture the more general writing style of an author. Below follows a high-level description of all implemented features.

Capital words: is defined as the number of uppercase word bigrams divided by number of word bigrams in total in a text.

Character n-grams: data-driven n-grams on character level (where a character n-gram is a contiguous sequence of n characters from a given sample of text).

Character upper-lower ratio: is defined as the number of upper-case characters divided by number of lower-case characters in a text.

Lexical diversity: defined as the amount of unique words divided by the total amount of words in a text.

Lix: a metric defined as:

$$(\frac{no.words}{no.sents}) + (\frac{no.long\_words}{no.words})$$

where $no.words$ is the number of words, $no.sents$ is the number of sentences, and $no.long\_words$ is the number of long words, defined to be all words that are longer than six characters. It is intended to be used as an approximation of readability and has in our initial experiments performed better than more traditional measures such as Flesch-Kincaid [8].

Masked character n-grams: work as the character n-grams, but with the difference that all characters between A and Z (uppercase as well as lowercase) are masked as a star (*). The idea is to have this feature focus on the effects of punctuation, spacing, diacritics, numbers, and other non-alphabetical symbols, as suggested in [3].

Part-of-speech (POS) tag n-grams: work in the same way as the word n-grams, but use POS tags as tokens rather than words. The POS tagging is language dependent and

relies on spaCy's[1] POS tagger.

Sentence length: for all sentences in a text, the mean and standard deviation are calculated to get the average sentence length and the variation of sentence lengths.

Shannon entropy: intended to capture the entropy of texts written by an author. It is defined as

$$H = -\sum_{i=1}^{M} P_i \, log_2 \, P_i$$

where P_i is the probability of character number i appearing in the stream of characters of the message.

Word length: for all words in a text, the mean and standard deviation are calculated to get the average word length and the variation of word lengths.

Word length distribution: a vector representing the raw counts of word lengths up to 16 characters, divided by the total amount of words in a text.

Word n-grams: data-driven n-grams on word level (where a word n-gram is a contiguous sequence of n words from a given sample of text).

5 Submitted Systems

Our research institute FOI has contributed with two system submissions to the PAN 2019 authorship attribution challenge. These submissions are here referred to as **isbister19** and **johansson19**, respectively. Since the submitted solutions have much in common we describe them jointly, but in practice we have performed much of the feature selection, experimentation, etc. more or less independent of each other. This means that none of the submitted systems utilize all of the features described in last section. Instead, they use (partially overlapping) subsets of these features. A detailed description of which features that are used in which system is given in Tables 1 and 2.

As previously described, our implemented systems build upon the baseline-SVM classifier, although it has been extended with a lot of other features. We have experimented with several other types of standard classifiers, but the linear SVM classifier performed consistently better than standard alternatives such as random forest and AdaBoost classifiers. However, it is important to note that the SVM classifier performed much better when using a one-vs-all regime, training as many binary classifiers as there are candidate authors. In the implemented classifier there is also a reject option, assigning an unknown document to the <UNK> class when the difference of the top two

[1] https://spacy.io/

Table 1. System overview - **isbister19**

Workflow	
Features	**Parameters**
Lix	
CharUpperLowerRatio	
CountWordCaps	
avg_sen_len	
std_sen_len	
lex_diversity	
avg_word_len	
std_word_len	
shannon_entropy	
word_sizes	
word_ngrams	weighting=smooth_idf, range(1, 3), min_df=2, lower=True
char_ngrams	weighting=smooth_idf, range(1, 3), min_df=25, lower=True
binary_char_ngrams	weighting=binary, range=(1, 4)
Transformation	
MaxAbsScaler	
Classifier	
LinearSVM (One-vs-all)	C=1.0

candidates is less than a pre-specified threshold (set to 0.1 in both systems). Initial experiments indicated that the classifiers perform better with this kind of solution rather than to assign <UNK> when the most likely class probability is less than some predefined threshold. However, systematic searches for good threshold settings have not been undertaken, so it is likely that classification performance can be increased by fine-tuning this threshold.

In case of the **isbister19** submission, the whole corpus for each problem (including the texts from the unknown authors as well) were used to create the vocabulary for the data-driven n-gram representations. A small increase of performance could be gained when using the vocabulary also from the unknown authors. Grid search has been used for both submitted systems in order to find good parameters for the n-grams. In both systems we have concatenated all used features into a single feature vector. This vector has been transformed by scaling each feature by its maximum absolute value. Somewhat surprisingly, the choice of scaling method had a huge impact on the predictive performance, as other standard scaling methods performed much worse.

6 Results

When developing our submissions, we decided on which features to include, how to scale the data, which classifier to use, etc. by validating different configurations on the development corpus being part of the PAN 2019 cross-domain authorship attribution dataset. The results for the submitted systems, as well as two provided baselines to

Table 2. System overview - **johansson19**

Workflow	
Features	**Parameters**
Lix	
CharUpperLowerRatio	
CountWordCaps	
avg_sen_len	
std_sen_len	
avg_word_len	
std_word_len	
word_sizes	
word_ngrams	weighting=smooth_idf, range(1, 3), min_df=3, lower=True
char_ngrams	weighting=smooth_idf, range(1, 3), min_df=3, lower=True
POS_ngrams	weighting=smooth_idf, range(1, 2), min_df=2, lower=True
Masked ngrams	weighting=smooth_idf, range(1, 2), min_df=2, lower=True
Transformation	
MaxAbsScaler	
Classifier	
LinearSVM (One-vs-all)	C=2.0

Table 3. Evaluation results (macro F1 scores)

Submission	**Train-dataset 1**	**Test-dataset 1**	**Test-dataset 2**
isbister19	0.641	0.607	0.622
johansson19	0.623	0.610	0.616
PAN2019-svm	0.576	xx	xx
PAN2019-compressor	0.556	xx	xx

which we have compared our results, as reported by the PAN 2019 evaluation script are shown in the first column of Table 3. These results are, however, not very reliable as various grid searches have been performed in order to find features and classifiers that perform well on this specific evaluation.

According to our preliminary analysis there was a clear difference in difficulty for different problem instances on this development corpus, but we could not see any distinct trend in that our trained classifiers performed better for some languages than others.

A more reliable estimate of how well the submitted systems can be expected to perform on unseen data is given in the two last columns of Table 3. These are the results obtained when submitting the systems to TIRA and evaluating them on previously unseen data. **Test-dataset 1** is the data that was used to evaluate the early bird submissions, while the **Test-dataset 2** is the larger dataset used to evaluate the final submissions.

7 Conclusions

In this paper we have presented our submitted systems for the cross-domain authorship attribution task at PAN 2019. The final submitted systems can be seen as extensions of the PAN-CLEF 2019 baseline-SVM system, to which we have added a large amount of hand-crafted stylometric features. The submitted systems have achieved overall scores of approximately 0.62 macro F1 on the final TIRA test set. They have also consistently outperformed the baseline implementations to which they have been compared.

As future work, we would like to contrast this type of models with more modern pre-trained deep-learning architectures such as language models implemented as stacked Long Short-Term Memories (LSTMs) [4] or Transformers [14], which are fine-tuned on the specific problem instances at hand. However, since such approaches would take a very long time to run on TIRA, given the current specifications of the virtual machines, such an evaluation would require the PAN 2019 cross-domain authorship attribution evaluation corpus (or similar datasets) to be released for such an evaluation to be feasible.

Acknowledgments

This work was supported by the R&D programme of the Swedish Armed Forces.

References

1. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26(2), 7:1–7:29 (Apr 2008)
2. Ashcroft, M., Johansson, F., Kaati, L., Shrestha, A.: Multi-domain alias matching using machine learning. In: 2016 Third European Network Intelligence Conference (ENIC). pp. 77–84 (Sep 2016)
3. Custódio, J., Paraboni, I.: EACH-USP Ensemble Cross-Domain Authorship Attribution—Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France. CEUR-WS.org (Sep 2018)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997), http://dx.doi.org/10.1162/neco.1997.9.8.1735
5. Johansson, F., Kaati, L., Shrestha, A.: Detecting multiple aliases in social media. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1004–1011. ASONAM '13, ACM, New York, NY, USA (2013)
6. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3), 233–334 (Dec 2006), http://dx.doi.org/10.1561/1500000005
7. Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)
8. Kincaid, J.P., Jr., R.P.F., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Institute for Simulation and Training, University of Central Florida (Jan 1975)

9. Layton, R., Watters, P., Dazeley, R.: Authorship attribution for twitter in 140 characters or less. In: Proceedings of the 2010 Second Cybercrime and Trustworthy Computing Workshop. pp. 1–8. CTC '10, IEEE Computer Society, Washington, DC, USA (2010)
10. Narayanan, A., Paskov, H., Gong, N.Z., Bethencourt, J., Stefanov, E., Shin, E.C.R., Song, D.: On the feasibility of internet-scale author identification. In: 2012 IEEE Symposium on Security and Privacy. pp. 300–314 (May 2012)
11. Pillay, S.R., Solorio, T.: Authorship attribution of web forum posts. In: 2010 eCrime Researchers Summit. pp. 1–7 (Oct 2010)
12. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
13. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60(3), 538–556 (Mar 2009), https://doi.org/10.1002/asi.v60:3
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017), https://arxiv.org/pdf/1706.03762.pdf
15. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics. SIGMOD Rec. 30(4), 55–64 (Dec 2001)