Stylometry Based Article Classification and Paper Fingerprinting

Ilija Simić
Graz University of Technology
Graz, Austria

Mauro Jurada
Graz University of Technology
Graz, Austria

Abstract

This paper aims to identify the potential of a visual representation of features related to authorship attribution, as well as evaluate which of the features would be most useful. Machine learning article classification was used for differentiating authorship profiles, which in turn provided us with an evaluation of stylometric features.

CCS Concepts: • **Information systems** → Content analysis and feature selection.

Keywords: Authorship attribution, machine learning, literature fingerprinting, high content similarity

1 Introduction

Two tasks based on stylometry were addressed, authorship attribution and profiles. Stylometry represents a study of linguistic style, which is used in authorship attribution and plagiarism check. Usually, there are two types of news sources: *high-quality press* and *yellow press* (also known as *tabloids* or *yellow journalism*). Since the yellow press is usually related to treating news in an unprofessional or unethical manner, it would be a good idea to have an automatic identifier whether the news source is a high-quality one or the news come from a tabloid. In our first task, we tackle this problem by looking only on the text of article, specifically its stylometry features and by using machine learning techniques we try to classify it to the right group. Besides classifying articles, we also attempt to create a visual representation of each text in the form of a fingerprint. We will be analysing the datasets and proposing a way to algorithmically identify individual authors/news sources. Both of these tasks are done in two phases, where first phase represent extraction of numerical features and second phase represent either classification or creation of visual representation.

2 Related Work

Inspiration for the fingerprinting direction we took was found in the [Keim and Oelke 2007] paper about literature fingerprinting. They visually described books through authorship attribution. In short, they attempted to find a distinct author style mostly through the average length of the sentences or through another stylometric feature. This paper aims to expand on theirs by showing how a fingerprint with multiple features is more suitable for authorship attribution.

From [Lex et al. 2010b] we draw most of the inspiration for article classification task. In [Lex et al. 2010b] authors assess objectivity in online news media. In this paper, the authors conducted three experiments. Firstly, they used the standard *bag-of-words* model for classifying articles into two categories - *High Quality*(perceived as objective) and *Yellow Press*(perceived as subjective). Here, authors noticed that classification performs better when stopwords are not removed and no stemming is performed. Secondly, they inspected the topic dependency of *bag-of-words* features in a *cross-domain* experiment. In this experiment, authors noticed that the *bag-of-words* model implicitly learn topics, due to significant accuracy drop for different topics. Thirdly, they investigated the applicability of *stylometric features* for objectivity classification. Even though the general accuracy for the same domain is a bit lower than the standard *bag-of-words* model, there is no significant drop for the *cross-domain* so *stylometric features* outperform the *word-based* approaches with respect to generalizability. *Stylometric features* are introduced in [Lex et al. 2010a] and they include punctuation distribution, emoticons, words per sentence, characters per sentence, noun-verb groups per sentence, the average number of unique part-of-speech tags per sentence, the ratio of lower to upper case characters, the word length distribution, the adjective rate, and the adverb rate.

From [Kern et al. 2011] we took ideas for two stage approach and for trying ensemble based classifier. In this paper authors, for the needs of PAN 2011 autorship identification challenge have developed two stage approach for identification of author. In first stage they generated multiple set of features and in the second phase train multiple classification models for each author based on these features. After that authors have combined these models to assign the author of email.

We had also intensively used [jpotts18 2020], a python library for calculating stylometric features, but before the usage we had to make it compatible with python 3.

3 Datasets description

Both datasets were manually collected and labelled.

3.1 Newspaper Articles

This dataset is created by collecting articles on group of topics, but from different sources inside the same topic. There are five topics:

1. Articles about first historical meeting of Donald Trump and Kim Jong Un
2. Articles about Kanye West announcing his presidential campaign
3. Articles about Tom Hanks becoming Greek citizen
4. Articles about Lewis Hamilton sharing fake news
5. Articles about publishing the book about Meghan Merkle and Prince Harry

Each of these groups has two articles from high-quality press and two from yellow press, except in second group where we have 4 of each group. Articles from *The Guardian*, *BBC News*, *Reuters*, *The Washington Post* and *Deutsche Welle* are labeled as high-quality, while articles from *The Sun*, *The Daily Mail*, *The Daily Mirror*, *Irish Post* and *The Daily Express* are labeled as yellow press. In total there is 28 articles, 14 classified as high-quality press and 14 classified as yellow press.

3.2 Scientific Papers

A collection of 26 scientific journals within the umbrella publisher Elsevier were gathered. To make them as comparable as possible, all of them are about *green roofs*. Handpicked so that the overall theme and contents are as similar as they can be without trying to.

Only parts of the papers were taken - following the same format. Usually split into four distinct categories all of them consist of:

1. Problem description
2. Definition/differentiation
3. Benefits
4. Downfalls

These four categories are very interesting for our purpose since even though the conclusion and the specific area of focus for each paper can be totally different, they all start with these sections/categories and then build on top of them. This allows for all the extracted texts to maintain full coherence, share plenty of keywords, define the same terms, and still be over 500 words each.

4 Feature Extraction

In this section, we will describe calculation of stylometric features and what they represent. We used stylometric features because they are topic independent, which is important due to different domains of articles. Before calculating features on articles, we had firstly needed to remove quotes, as quotes are not represent of writer style. Only feature that is calculated on original text is *number-of-quotes*. First group of features represents traditional stylometric features. They are calculated using stylomtetry library. It includes: *alpha-chars-ratio, digit-chars-ratio, upper-chars-ratio, white-chars-ratio, number-of-words, size-of-vocabulary, type-token-ratio, hapax-legomena, hapax-dislegomena, mean-word-length, mean-sentence-char-length, mean-sentence-word-length, mean-paragraph-length, exclamation-mark-rate, question-mark-rate.* Second group of features represents POS based features. This group comprise of *adverbs-rate, adjectives-rate, nouns-rate, prepositions-rate, conjunctions-rate, verb-rate, adjectivs-adverbs-rate, named-entities-count.* These features are calculated using NLTK's part of speech tagger, apart from *named-entities-count* which was calculated by Spacy.

Alpha-chars-ratio represents the rate of alphabet characters in the text.

Digit-chars-ratio represents the ratio of digits in the text.

Upper-chars-ratio represents the ratio of upper case letters in the text.

White-chars-ratio represents the ratio of the white-space characters in the text.

Number-of-words represents the total number of words in the text.

Size-of-vocubalary is the total number of different words used in the text.

Type-token-ratio represents the ratio of the former two values.

Hapax-legomena represents number of words occurring once, while *hapax-dislegomena* represents number of words occurring twice.

Mean-word-length represents the average length of word.

Mean-sentence-char-length represents the average length of sentence per number of characters, while *mean-sentence-word-length* represents the average length of sentence per number of words.

Mean-paragraph-length represents the average length of paragraph per number of words.

Exclamation-mark-rate represents the rate of the exclamation mark(!), while the *question-mark-rate* represents the rate of the question mark(?).

Adverbs-rate represents the ratio of adverbs in the text. *adjectives-rate* represents the ratio of adjectives in the text, while *adjectives-adverbs-rate* represents the ratio of the sum of these two type of words in the text.

Nouns-rate, prepositions-rate, conjunctions-rate, verb-rate represents ratio of nouns, prepositions, conjunctions and verbs in the text.

Named-entites-count represents the ratio of named entities in the text.

5 Feature Evaluation and Visualisation

With a plethora of features to choose, it is easy to fall into the trap of overfitting. It is important to pick only a few features. Ones that might uniquely describe a specific author/source or even fully identify the author from a reference corpus of multiple authors.

5.1 Machine Learning

In this section we will address the problem of classifying the article in the correct group. Firstly, we had to divide

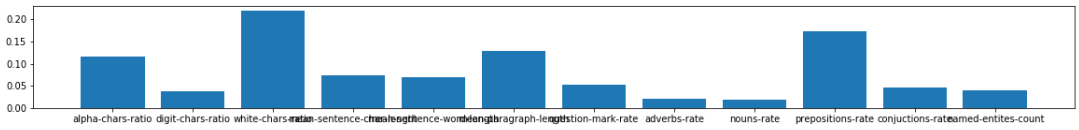

Figure 1. Bar plot showing features importance

Accuracy	0.9
Precision	1.0
Recall	0.88
F1 Score	0.889

Table 1. Results of classifier on test set

the dataset in a *training* and a *test* set. Since our classes were equally distributed, dataset was randomly split into subsets, where test set size is 35% of total set. This means that training set had 18 samples, while the test set had 10 samples. We tried different classifiers, such as AdaBoost, Random Forest, Multi-layer Perceptron and Extremely Randomized Trees. Extremely Randomized Trees showed best performance and it is used for determining the most important features. Ten most important features are *alpha-chars-ratio, digit-chars-ratio, white-chars-ratio, mean-sentence-char-length, mean-sentence-word-length, mean-paragraph-length, question-mark-rate, adverbs-rate, nouns-rate, prepositions-rate, conjunctions-rate, named-entities-count*. After choosing the most important features, we retrained the classifier using only this subset of features. The results we got are shown in table 1. The feature importance is shown on figure 1. We can see that the most significant feature is *white-chars-ratio*, and by looking at its distribution (figure 2) we see that yellow press usually have slightly higher values than high quality press. This could be explained by the fact yellow press usually have more shorter paragraphs instead of few longer ones so that increases number of white space characters. Also, from the list we see that features from the [Lex et al. 2010b] such as *adverbs-rate* and *mean-sentence-word-length* were also included as important ones.

5.2 Fingerprinting

Selecting the twelve most highly impactful features we can create a visual fingerprint of each text. Its aim is to clearly present the texts uniqueness through its attributes.

These already previously selected attributes are always shown in the same order (Figure 3). The two horizontal lines represent the mean of these attributes throughout the whole dataset. Left and right positioning of an attribute shows if the value of it is lower or higher than the mean. This distance is scaled for the sake of readability. The connecting lines aim to make the shape of the fingerprint memorable and distinguishable.

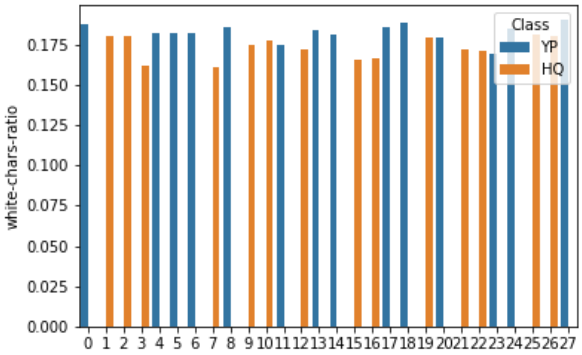

Figure 2. White-chars-ratio distribution bar

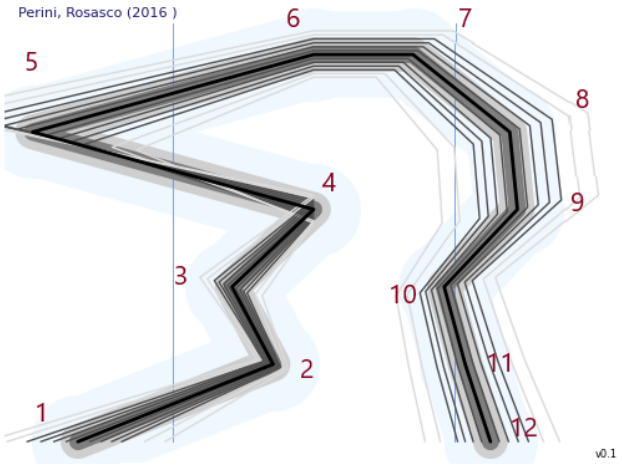

Figure 3. Annotated fingerprint. 1: *Alpha-chars-ratio*, 2: *Digit-chars-ratio*, 3: *White-chars-ratio*, 4: *Size-of-vocabulary*, 5: *Type-token-ratio*, 6: *Hapax-legomena*, 7: *Mean-word-length*, 8: *Mean-sentence-char-length*, 9: *Mean-sentence-word-length*, 10: *Adverbs-rate*, 11: *Nouns-rate*, 12: *Prepositions-rate*

In Figure 4 we can see that the text has higher emphasis on *alpha-chars-ratio*, *mean-word-length* as well as a lower *digit-to-chars-ratio*. On the other hand, the text from figure 5 seems to be fairly average except it has lower *Mean-sentence-char-length* and *Mean-sentence-word-length* than the others.

While creating a fingerprint for individual texts, it also makes sense to visually show how comparable they are. In

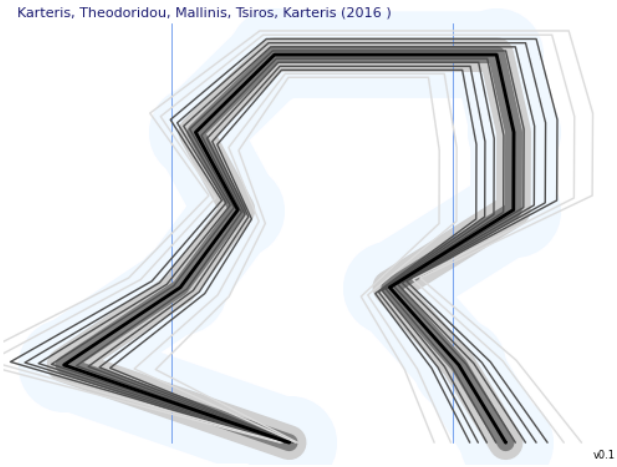

Figure 4. Fingerprint of a scientific paper on green roofs by Karteris, Theodoridou, Mallinis, Tsiros and Karteris (2016).

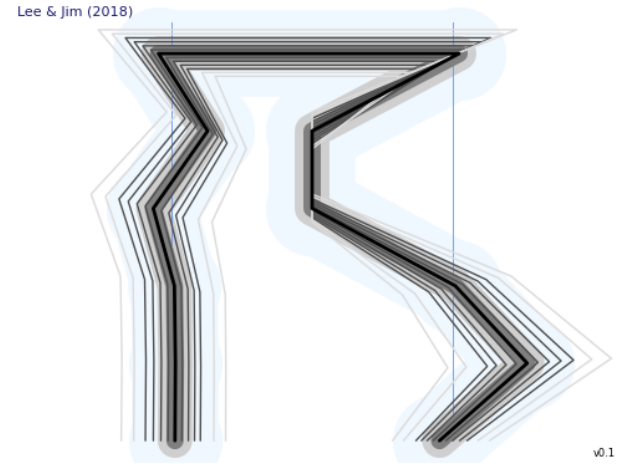

Figure 5. Fingerprint of a scientific paper on green roofs by Lee & Jim (2018).

figure 6 we can immediately see how the two texts compare. And should more texts from the same authors be collected and compared to another, their difference should be obvious. This would be done by cherry picking certain attributes from one author that remain completely consistent throughout their writing.

This fingerprinting tool was made to be very flexible and can handle any text or attributes - provided it is in the correct format. Expected format is a *CSV* table with a name column and ten or twelve attribute columns. This number of attributes was decided on because fingerprints are most visually appealing when showing five or six attributes on each side. These columns are completely free to be picked based on any criteria; currently the attributes are decided based

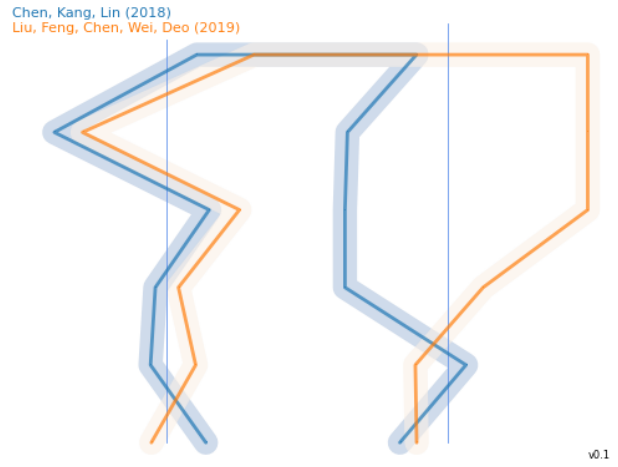

Figure 6. Two fingerprints compared to show differences in the texts.

on which attribute had the most impact in the 5.1 Machine Learning section.

6 Conclusion

We used stylometric features to determine quality of news source with machine learning to determine which of these stylometric features were more relevant for prediction. Even though dataset was quite small, ensemble based classifier achieved 90% accuracy on assigning correcting quality of press. It showed us that stylometric features can be used for determining the quality of the article. Additionally, from the feature importance we found features that are not only relevant for prediction, but also explainable.

Visual representation of features in a fingerprint format is much easier to read. Provided that a reasonably sized dataset exists, comparisons between fingerprints can be used for detecting if someone really wrote the text they claim they did. Flexibility of the tool allows for interchanging of the features observed. Perhaps this might sound counter-intuitive because if someone does not know which features are shown the fingerprint might be useless. Even so, it is important to first find the most heavily impactful features to show. Over time, with more data, choice of features used for fingerprint might, and should, be streamlined. Features might be grouped and the meaning might be immediately obvious to everyone (e.g., put some features that show quality of writing on one side so that if lines are to the left the text might be considered of poorer quality).

References

jpotts18. 2020. Python stylometry library. https://github.com/jpotts18/stylometry. Accessed: 2020-08-30.

Daniel Keim and Daniela Oelke. 2007. Literature Fingerprinting: A New Method for Visual Literary Analysis. *VAST IEEE Symposium on Visual*

Analytics Science and Technology 2007, Proceedings (10 2007), 115 – 122. https://doi.org/10.1109/VAST.2007.4389004

Roman Kern, Christin Seifert, Mario Zechner, and Michael Granitzer. 2011. Vote/Veto Meta-Classifier for Authorship Identification.

Elisabeth Lex, Michael Granitzer, Markus Muhr, and Andreas Juffinger. 2010a. Stylometric features for emotion level classification in news related blogs. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*. 132–133.

Elisabeth Lex, Andreas Juffinger, and Michael Granitzer. 2010b. Objectivity Classification in Online Media. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (Toronto, Ontario, Canada) *(HT '10)*. Association for Computing Machinery, New York, NY, USA, 293–294. https://doi.org/10.1145/1810617.1810681