Similarity Ranking using Handcrafted Stylometric Traits in a Swedish Context

Johan Fernquist, Björn Pelzer, Lukas Lundmark, Lisa Kaati, Fredrik Johansson
Department of Defence Technology
Swedish Defence Research Agency
Kista, Sweden
firstname.lastname@foi.se

***Abstract*—In this paper we introduce a new type of handcrafted textual features called *stylometric traits*, used to create a stylistic writeprint of an author's writing style. These can be divided into four categories: word variations, abbreviations, internet jargon and numbers. A *similarity ranking* method is developed for ranking users' social media accounts based on how similar their writeprints are. We experiment with both vector distance metrics and machine learning-based class probabilities to measure similarity. The best performance is achieved using stylometric traits combined with the Jensen-Shannon distance metric, outperforming traditional stylometric features used in previous research.**

***Index Terms*—Alias matching, similarity ranking, stylometric traits**

I. Introduction

Authorship analysis can be defined as the study of the linguistic style of a written text. Usually, authorship analysis is used to either find authors that have similar writing style or to attribute the authorship of an anonymous text to a known author. The most common approach in authorship analysis is to use a set of stylometric features and create a (potentially unique) writing profile of an author (often referred to as a *writeprint* [1]). Comparing authors based on their stylometric profiles is sometimes referred to as *similarity detection* [1]. Similarity detection is within a social media context a central component for *alias matching* [4], [9], i.e., the linkage of social media accounts belonging to the same user.

We aim at developing a method that is able to rank each author in a finite set of authors based on how similar their writeprints are compared to the writeprint of an anonymous author. We refer to this problem as *similarity ranking*. The long-term goal is to be able to use similarity ranking in a law enforcement context to identify a set of candidate social media user accounts which are more likely to belong to an anonymous author.

The most common approach to create a stylistic profile of an author is to use stylometric features such as function words and the vocabulary richness of the language. Many stylometric features are language independent and can therefore be transferred and applied in several different languages. We present a new set of features that we denote *stylometric traits*. The stylometric traits are a set of handcrafted features that can be used to create a stylistic writeprint, allowing for comparing the similarity of two or more texts. The stylometric traits we propose are based on an in-depth analysis of the writing of internet users on a large Swedish discussion forum. Most existing research has been focusing on English, while Swedish is an example of smaller languages for which significantly less authorship analysis-related research has been carried out. We use the stylometric traits to detect forum users with similar writing style and test our approach on three different data sets with varying difficulty, where the most challenging is an in the wild dataset. Our experimental results show that using stylometric traits to rank users with similar writing style outperforms more traditional stylometric features. We also show that the Jensen-Shannon distance metric consistently perform better than other distance metrics and alternative approaches for measuring similarity based on the extracted stylometric traits.

This paper is outlined as follows. Section II describes previous research on the subject of authorship analysis. In Section III we present a set of handcrafted stylometric traits that are based on an in-depth analysis of the writing style of Swedish internet users. In Section IV we describe how the stylometric traits can be used to rank authors based on the similarity of their writing style. Section V describe our experimental setup and the different datasets we have used in our well-controlled experiments. The results of these are presented in Section VI. Finally, in Section VII we present conclusions and ideas for future work.

II. Related Work

The use of stylometric features for the purpose of authorship attribution has been a subject of research for many years. Early works such as [7] and [14] had their origins in determining authorship of disputed works of literature, and they predate social media and digital texts by decades. The rise of the internet has enabled millions of users to publish texts for a global audience in blogs, forums or social media, at little to no cost, and anonymously. This in turn has provided researchers with access to vast amounts of texts, some of which may be of forensic interest, for example anonymous online death threats. The number of accessible texts presents ample opportunity to study automated approaches. At the same time these texts come with their own challenges, in particular their brevity. Where earlier research investigated books or at least substantial pamphlets like the Unabomber

Manifesto [12], social media texts are typically much shorter. For example, the average post on Reddit has merely 30 words and 180 characters, and Twitter has traditionally even imposed a maximum length of 140 characters - this was increased to 280 in 2017, yet in practice the average tweet is even substantially shorter at approximately 11 words or 67 characters.[1] This means that automated methods may have little data to work with per individual user.

Much of the early work on authorship analysis-related research has been scattered among several research fields. Stamatatos et al. [21] have made an excellent work of compiling an overview of such approaches. In later years, authorship attribution and authorship verification have become a mainstay of the annual PAN competition, with the event featuring one such task every year since 2011 [3], and future tasks already being planned. PAN is relevant as many of the more promising authorship attribution and verification methods are evaluated in these competitions. PAN provides extensive training datasets and is a good way to find out how different methods compare to each other in terms of generalization capability to new unseen test datasets. Therefore, PAN gives an idea of how well these methods work in a more real-world setting. In 2020 most of the competing systems employed various machine learning-approaches [13], including the three leaders, although they differ in the specific features: the winning system by Boenninghoff et al. [5] relies on deep learning to discover and learn features automatically, while the runners-up use stylometric features [8], [25]. The former approach has the advantage of being able to exploit powerful non-obvious features, but at the same time the decision of such a system is difficult to interpret for a human investigator, and the method risks overfitting to the specific topics being represented in the training data. The competition setup differs from the social media situation in critical aspects: the texts have an average length of 21,000 characters, providing orders of magnitude more data than forum posts or tweets, and the set of authors is closed, i.e. the test set contains only authors already encountered in the training set. Also, the competition text samples have been obtained from fan fiction and thus do not account for adversarial situations where authors attempt to change their style in order to mask their identities.

There have been numerous attempts at tackling the specific challenges of authorship analysis in social media, and Rocha et al. [17] provide a comprehensive overview. As texts from a given author may be scarce, short, and scattered across topics and media, much effort goes into exploring topic-independent features that remain robust over limited data, such as Sapkota et al. [18] who work with texts with at least 600 words. Greeting patterns, signatures, white-space and punctuation are used for authorship attribution on emails in [6] and [22]. Success on actual Twitter-scale messages is mixed, though. For example, Schwartz et al. [19] achieve 60 % accuracy on closed sets of 50 authors with at least 200 Tweets each. Andrews and Bishop [2] work with considerably more open sets from Reddit and Twitter - given users are compared against 111,000 and 170,000 targets, respectively. Using both stylometric features and some meta-data (time of posting and category), the method achieves an $R@32$ of 0.3 for Twitter and an $R@8$ of 0.65 for Reddit.[2]

[1]Based on random samples of 1 million posts each for Reddit and Twitter.

III. Stylometric Traits

The stylometric traits that we use are variations of commonly used and topic independent words, phrases and abbreviations that are specific to an individual's writing style. Such traits can for example be how someone uses abbreviations, how punctuation is used or how someone writes numbers. We use four categories of stylometric traits that we have identified based on in-depth studies of real forum posts. Table I shows the stylometric traits that we have identified, together with some Swedish examples and their English translations. Also included are English examples that exhibit the traits, as the English translations of the Swedish examples tend to lose this feature. In total, we have identified 260 stylometric traits that are used throughout our experiments. As will be seen, these stylometric traits can when combined be good indicators of authorship.

1) Word variations: Some words have a variety of forms, where different authors prefer to use different variants. This category covers the use of synonyms, where several words mean the same thing, and also alternative spellings of a particular word, such as *yes*/*yeah* or *impostor*/*imposter*. Reasons for preferring a certain word or spelling can for example be related to the author's age or origin.

2) Abbreviations: Abbreviations can be written in a variety of ways using punctuation, blank spaces or by just concatenating the initial letter of words without any punctuation or blank space in between. This is especially common for abbreviations that consist of several words such as *asap* (as soon as possible) or *eg* (example given).

3) Internet jargon: Internet jargon consists of abbreviations and slang words that are the result of an increased use of digital communication. Many of these words are phrases that are put together into one word, and the words are often spelled based on how they are pronounced rather than their formally correct spelling. An English example for this is *whatchadoin*.

4) Numbers: Some authors prefer to write numbers with letters rather than digits. In some languages such as Swedish, the most grammatically correct way to write numbers up to twelve (in free text) is with letters. The use of digits instead of letters can therefore very well contribute to an individual's unique writing style.

IV. Similarity Ranking with Stylometric Traits

Similarity ranking can be used when identifying users that have a non-public identity, e.g., users who are employing different usernames or handles on different social media platforms. Law enforcement agencies and intelligence analysts often have to deal with cases where an unidentified suspect is

[2]$R@k$: recall at the top-k ranked results

TABLE I
STYLOMETRIC TRAITS

Category	Examples of stylometric traits	English translation	English examples	Count
Word variations	denna, den här inte, icke, ikke, ej, inge de, dem, dom	this one not they	yes, yeah, yup meager, meagre	84
Abbreviations	till exempel, t.ex., t ex på grund av, p.g.a, p g a något, nått, nåt, ngt	for example because of something	e.g., eg MSc, M.Sc., MS	131
Internet jargong	är det, äre, ere vad fan, va fan, vafan vara, va	is it what the * to be	whatchadoin noob dunno	20
Numbers	två, 2 tio, 10 elva, 11	two ten eleven	two, 2 ten, 10 eleven, 11	25

known to have written one or several texts. To find out more about the suspect, analysts may want to identify social media accounts that are likely to belong to the suspect. In such cases it is common to conduct a manual in-depth analysis of the writing style, and to identify social media accounts that share the use of uncommon spelling errors, stylistic mannerisms, etc. with the texts known to have been written by the suspect. A drawback of such a manual analysis is that it scales badly, and it can only be used when very uncommon or unique language constructs have been used.

Instead of manually analysing the writing style of an author, various computer-based authorship analysis technologies can aid the analyst. For the given problem, authorship attribution techniques are in general a bad fit as they assume that the true author is among the set of candidate authors. In practice it is far from certain that there should be a match between the anonymous suspect's texts and any texts from user accounts it is being compared to. Instead, we treat it as a similarity ranking problem. The intended use of our proposed similarity ranking approach is to rank a set of social media users according to how similar their writing styles are to the texts written by the unidentified suspect. Such a comparison can be made in various ways, but we suggest the use of the stylometric traits presented in Section III. The idea is to automatically identify the top-n candidates with the most similar writing style and to present them to the analyst for further manual analysis. In this way, the analyst gets support with reducing the number of candidates, but is still in charge of investigating the best matches more closely.

To allow for comparison with previous work we have also included traditional stylometric features as described in [15] and [11]. The stylometric features that we use are listed in Table II. Each category of features are individually normalized so that the sum of all features in the category summarizes to 1. There are a total of 197 stylometric features.

TABLE II
STYLOMETRIC FEATURES

Category	Description Relative frequency of:	Count
Word lengths	words with 1-20 characters	20
Letters	a-ö (ignoring case)	29
Digits	0-9	10
Punctuation	characters . ? ! , ; : ( ) ” -	10
Stop words	stop words	114
Smileys	smileys :') :-) ;-) :P :D :X <3 :) ;) :@ :* :| :$ %)	14

V. METHOD

The similarity ranking problem requires the ability to achieve a measure of how similar two numerical vectors are. We have experimented with different machine learning models and vector similarity metrics to identify the best similarity measure for the purpose of similarity ranking. Each step is described in more detail below.

A. Experimental setup

In our experiments, we use data from discussion forums and therefore we will in the following use the terminology *user* when referring to an author of a text. Our overall goal is to rank a set of users $U = \{u_1, u_2, \ldots, u_n\}$ according to their similarity to a user x. The most similar user in U is ranked as number one, the next most similar as number two and so on.

To test different approaches of similarity ranking we use the following approach. First, we collect texts written by n users, resulting in n sets of texts, with the set u_i containing the texts written by the i-th user. Each such user u_i is then split into two separate users u_{ia} and u_{ib}, which gives us two sets of users $A = \{u_{1a}, u_{2a}, \ldots, u_{na}\}$ and $B = \{u_{1b}, u_{2b}, \ldots, u_{nb}\}$. In this paper we use different methods for splitting users, and these are described in Section V-B. This means that for every user $u_{ia} \in A$, there is a user $u_{ib} \in B$ that corresponds to the same original user u_i. For each user u_{ia} and u_{ib} a stylistic profile is created. One by one each user from $A \cup B$ is then selected as u_{sel} and compared to each other user in $(A \cup B) \setminus \{u_{\text{sel}}\}$ using the similarity measure of choice, and the users in $(A \cup B) \setminus \{u_{\text{sel}}\}$ are ranked according to their similarity to u_{sel}, from most similar to least. The reported accuracy is the fraction of times that the index of the selected user is found within the top-N rankings, i.e. if u_{ia} was selected, then u_{ib} needs to be found in the top-N, and vice versa. The approach of splitting users to solve the similarity ranking problem has been used before in [11] [10] and [20].

Example: we have 100 users. Each user is split into two new users, giving us two sets of users, $A = \{u_{1a}, u_{2a}, ..., u_{100a}\}$ and $B = \{u_{1b}, u_{2b}, , ..., u_{100b}\}$. We select every user from each set once, and compare that selected user to all other users. Hence, we compare u_{1a} to all other 199 users and obtain a ranking of their similarity to u_{1a}. To calculate the accuracy, we investigate where user u_{1b} (same index, originates from the same author) is ranked. If we calculate top-1 accuracy and u_{1b} is ranked as the most similar user to u_{1a}, this is considered a match. This is then done for all other 199 users, and each time the same index is found as the most similar user, it counts as a match. When all comparisons have been made, the accuracy is calculated as the number of matches divided by the total number of users after splitting (in this case 200).

To obtain a ranking of similarity between users and their stylistic profiles we use two different ways of measuring similarity: vector distance metrics and machine learning-based class probabilities.

1) Distance metrics: There are several different metrics for calculating the distance between two vectors. The vector distance metrics we have used are Cosine similarity, Euclidean distance, Bray Curtis, Canberra, City Block, Correlation and Jensen-Shannon. All distance metrics are calculated using Python's SciPy library [23], and they are further described in [24].

2) Machine learning approach: For the machine learning approach we build binary classifiers for the problem to classify whether two users are the same or not. This is done by creating a stylistic profile for each user, calculating the absolute difference between two profiles and then using the difference vector as input for the machine learning classifier. The classifier can then be used for new pairs of users, but instead of classifying them as not same user (class 0) or same user (class 1) we can extract the prediction probability. The prediction probability can then be used as the metric for ranking which user is most similar. The higher the prediction probability is between two users, the more likely they are to be the same.

The machine learning models are based on the stylometric traits and the stylometric features. To create training features for the machine learning models, we calculated difference vectors for pairs of vectors, $[|v_1[1] - v_2[1]|, |v_1[2] - v_2[2]|, \ldots, |v_1[n] - v_2[n]|]$ where v_1 and v_2 are the vectors and n the number of elements.

For every user u_{ia} in the training set, we have calculated two training examples. The positive training example is calculated by taking the difference vector between u_{ia} and u_{jb}, where $i = j$. The negative (i.e., not same user) training examples have been calculated by taking the difference vector of u_{ia} and u_{jb}, where $i \neq j$. We used two different machine learning algorithms, namely support vector machine (SVM) and logistic regression (LR).

All machine learning algorithms are implemented using Scikit-learn [16], and the algorithms are further described in Scikit-learn's documentation.

B. Data

We want to investigate how our stylometric traits perform on different types of data, so that we can detect strengths and flaws with the approach. Previous research shows that different strategies for splitting a user can result in varying problem difficulty. In [26] it is shown that splitting a user by randomly assigning posts yields an easier problem (higher accuracy) compared to splitting users by chronology. We expect that this has to do with randomly split users being more likely to have texts in both splits that stem from the same topic or conversation. This motivates us to increase the difficulty (and realism) by more complicated splits. In the experiments for this paper we use three different datasets, where the two first are obtained using different splitting strategies. These three datasets are described further below.

1) Split users: The data in the *split users* dataset is collected from one of Sweden's largest online discussion forums. We will refer to this as *Forum 1*. The only requirement to be included is that the user should have written at least 1,000 posts on the forum. Each user u_i is then split into two users u_{ia} and u_{ib} by randomly assigning posts written by u_i to either u_{ia} and u_{ib} while making sure that each subset holds the same number of posts. There is no restriction regarding the posts having the same topic or not. This experimental setup is similar to what has been used in previous studies.

2) Different topics: To simulate a more realistic setting, we ensure that the texts written by users u_{ia} and u_{ib} are not about the same topics. The data is collected from the same Forum 1 as the *split users* dataset. To ensure topic diversity we select a random set of users who have written at least 100 posts in at least two different subforums. The subforums focus on a variety of different topics, and it is highly unlikely that discussions from two different subforums will be about the same topics. Each user u_i is split into two users u_{ia} and u_{ib} where each user contains posts that are from different subforums. There is no restriction that u_{ia} or u_{ib} have to hold the same number of posts.

3) Cross domain: For the *cross domain* dataset, we collect data written by the same user, but on different online forums. One of the forums is the same Forum 1 as used for the *split users* and *different topics* datasets. The other forum is another large Swedish online discussion forum, which we will refer to as *Forum 2*. First we extract all usernames that are present on both forums. From this pool of users we remove those that have written less than 2,000 characters on each forum (corresponding to approximately half an A4 page of text) and then sample 4,000 of these usernames. A potential problem is that usernames in general are not unique identifiers, some common usernames might be used on different forums by different persons. To decrease the risk of such cases we removed usernames that were not unique enough to assume that there was one single person behind both users. To identify such names for removal we did a manual annotation of the names where we applied several rules: A username was to be excluded if it was a real name, a family name or a name of a known (fictional) character. A username was also to be

TABLE III
SUMMARY OF EACH OF THE DATASETS

Dataset	Source	Split condition
Split users	Forum 1	Each post randomly assigned to either u_{ia} or u_{ib}
Different topics	Forum 1	Split by subforums to ensure there is no topical overlap between u_{ia} and u_{ib}.
Cross domain	Forum 1 & Forum 2	All posts from users with unique usernames present on both forums. u_{ia} contains all posts from Forum 1, and u_{ib} all posts from Forum 2.

excluded if it was considered to be a common, well known word with a regular spelling. The annotation was done by three persons who individually annotated all of the randomly selected 4,000 usernames occurring in both forums. We only kept usernames that all of the three annotators considered to be sufficiently unique. This yielded a total of 2,251 usernames. In the final dataset, each user u_i is split into two users u_{ia} and u_{ib} where u_{ia} contains all the user's posts from Forum 1, and u_{ib} all the user's posts from Forum 2. This dataset is interesting in that it gives a better idea of how well the proposed method works in the wild, rather than just simulating a realistic setting.

C. Experiments

1) Experiment 1 - Training Classifiers: To be able to use machine learning prediction probability as a measure of similarity, classifiers have to be trained. The machine learning models are trained with 31,027 users from the *different topics* dataset, yielding a total of 62,054 training examples. To get an idea of the performance of the different classifiers, 5-fold cross validation is performed.

2) Experiment 2 - The Best Approach: To determine the best approach for the similarity ranking problem we test our machine learning models and distance metrics on a dataset consisting of 1,000 users from the *different topics* dataset. There is no overlap of users between the data used for training the machine learning models and the data used here. We test all distance metrics and the machine learning models' prediction probability functions and calculate the top 1 and top 10 accuracy.

3) Experiment 3 - Similarity Ranking in the Wild : In this experiment we investigate how the suggested approach performs on data from the real world. We take those machine learning and distance metric approaches from experiment 2 which are performing the best and use them on real-world data consisting of 4,502 users from the *cross domain* dataset.

4) Experiment 4 - Increasing the N: In this experiment we investigate how the accuracy increases by an increasing N, i.e., the number of user accounts presented as potential matches. This experiment is conducted with the best models from experiment 2 on the real-world data from experiment 3. We increased the group of most similar users up to N=1,000.

5) Experiment 5 - Increasing Posts and Dataset Complexity: The aim with this experiment is to investigate how our method performs on data split in different ways and the importance of the amount of text data for the similarity ranking problem. For this experiment, we use all three different datasets: the *split users* dataset, the *different topics* dataset, and the *cross domain* dataset. To make it possible to compare results between datasets we select 1,000 users from each dataset, and increase the number of posts from 50 to 1,000.

VI. RESULTS

1) Experiment 1 - Training Classifiers: Performance for each model is shown in Table IV. For each classifier, the stylometric traits outperforms the stylometric features.

2) Experiment 2 - The Best Approach: Top 1 and top 10 accuracy for each feature type and method is shown in Table V.

The stylometric traits outperforms the stylometric features for all methods. The best accuracy (both top 1 and top 10) is obtained with the distance metric Jensen-Shannon where the correct user is found almost 83 % of the time and within top 10 around 91 % of the time. The best accuracy for the stylometric features was achieved when calculating similarity with SVM. In this case the stylometric traits still outperformed the stylometric features with 0.721 in accuracy compared to 0.633. The machine learning model which performs best is SVM. The best distance metric for stylometric features is Canberra, but both machine learning models outperform Canberra.

Based on these results, using stylometric traits with Jensen-Shannon distance seems to be the best choice for the similarity ranking problem.

3) Experiment 3 - Similarity Ranking in the Wild : The results when using the best models from experiment 2 on real-world data are shown in table VI. When using Jensen-Shannon on the real-world data, the achieved accuracy was 15.3 % for top 1 and 23.4 % for top 10 using stylometric traits. The performance was quite worse using stylometric features, resulting in the accuracy 2.1 % (top 1) and 4 % (top 10). For the SVM, the stylometric traits still outperformed the stylometric features with top 1 accuracy of 9 % compared to 4.6 %. These results are significantly worse than for experiment 2. A reasonable explanation for this is the larger amount of users we have used in this experiment, since having more users to compare with yields a harder problem. In experiment 2, the users we investigate are split by topics, but every true pair of texts originates from the same user. In this experiment, the users come from two different forums and the splits from experiment 2 might result in a much easier problem. The amount of text for each user might also make the problem harder. Another possible reason might be that the stylometric traits are not stable between different domains.

4) Experiment 4 - Increasing N: Figure 1 shows how the top-N accuracy changes with increasing N. As can be noted, the accuracy shows logarithmic growth, and by increasing N, the highest accuracy boost is achieved for small Ns. For low N the stylometric traits outperforms the stylometric features, but for N larger than 700, the stylometric features with SVM starts to perform best.

TABLE IV
PERFORMANCE MEASUREMENTS FOR 5-FOLD CROSS VALIDATION FOR DIFFERENT FEATURES AND CLASSIFIER ALGORITHMS

		Accuracy	Precision	Recall	F1
SVM	**Traits**	0.940	0.940	0.940	0.940
	Stylometric	0.936	0.935	0.938	0.936
LR	**Traits**	0.918	0.906	0.934	0.920
	Stylometric	0.910	0.900	0.921	0.910

TABLE V
TOP 1 AND TOP 10 ACCURACY FOR 1,000 USERS FROM THE DIFFERENT TOPICS DATASET

	Top 1		**Top 10**	
	Traits	**Stylometric**	**Traits**	**Stylometric**
Cosine similarity	0.563	0.132	0.739	0.262
Euclidean distance	0.558	0.127	0.736	0.256
Bray Curtis	0.750	0.280	0.882	0.439
Canberra	0.704	0.500	0.853	0.687
City block	0.750	0.283	0.882	0.441
Correlation	0.559	0.131	0.736	0.260
Jensen-Shannon	**0.827**	0.252	**0.913**	0.374
SVM	0.721	0.633	0.880	0.837
Logistic regression	0.600	0.518	0.809	0.745

TABLE VI
TOP 1 AND TOP 10 ACCURACY FOR 4,502 USERS FROM THE CROSS DOMAIN DATASET ON THE BEST MODELS FROM EXPERIMENT 2

	Top 1		**Top 10**	
	Traits	**Stylometric**	**Traits**	**Stylometric**
Jensen-Shannon	0.153	0.021	0.234	0.040
SVM	0.090	0.046	0.173	0.131

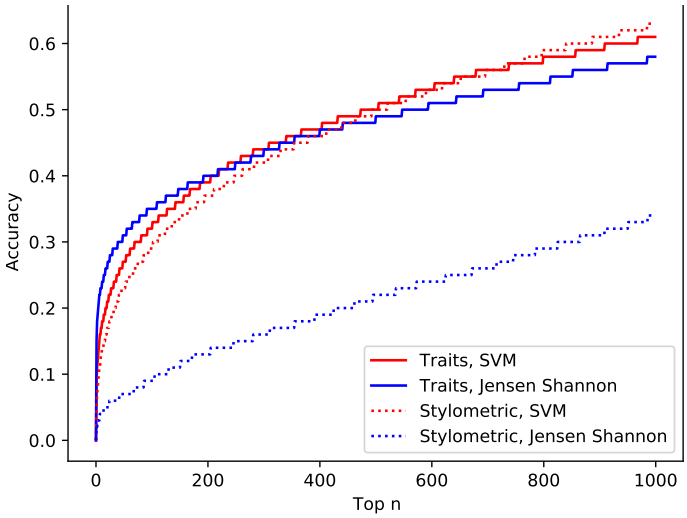

Fig. 1. Accuracy for experiment 4 with increasing top n ranking for the *cross domain* dataset

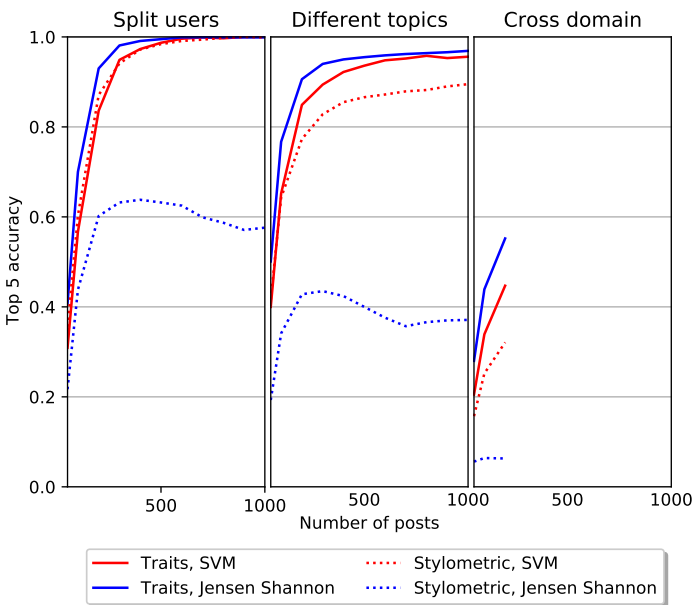

Fig. 2. Top 1 accuracy for experiment 5

5) Experiment 5 - Increasing Posts and Dataset Complexity: In Figure 2, the results for the three different datasets are shown. Due to the lack of a large amount of text data for the *cross domain* dataset, the experiments on the cross domain data was executed in fewer steps of increasing posts.

It is clear that increasing the number of posts contributes to increasing the accuracy. For the *split users* dataset, the accuracy converges towards 1 with an increasing number of posts. The accuracy for the *different topics* dataset converges around 0.93 for our best model. Due to the lack of data it is not possible to detect any convergence accuracy for the *cross domain* dataset.

The results show that the similarity ranking problem is much easier on the *split users* dataset and that the cross domain data is more challenging. As seen before, stylometric traits with Jensen-Shannon outperform the other methods, but for

stylometric features, the SVM provides a higher accuracy compared to Jensen-Shannon.

VII. Conclusions and Future Work

For the experiments conducted in this paper, it is obvious that the stylometric traits outperform the traditional stylometric features. We can conclude that in a case where an analyst wants to find social media accounts of a suspect, manually investigating just a few more accounts can give a large pay off since investigating just a few more accounts will increase the likelihood that the suspect's account is included. Our experiments show that the amount of text have a significant impact on the accuracy, where more text yield a higher accuracy. We further show that performing similarity ranking on real-world data from multiple sources is complex problem compared to when performing similarity ranking on datasets with users from one source.

For future research we plan to translate the stylometric traits to other languages. This will enable us to evaluate the traits on established benchmark datasets, comparing them to other features and approaches for the similarity ranking problem. We will also use the stylometric traits on other problems, such as authorship verification and authorship attribution.

References

[1] A. Abbasi and H. Chen. 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] N. Andrews and M. Bishop. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[3] S. Argamon and P. Juola. Overview of the International Authorship Identification Competition at PAN-2011. In V. Petras, P. Forner, and P. Clough, editors, *Notebook Papers of CLEF 2011 Labs and Workshops, 19-22 September, Amsterdam, Netherlands*. CEUR-WS.org, Sept. 2011.

[4] M. Ashcroft, F. Johansson, L. Kaati, and A. Shrestha. Multi-domain alias matching using machine learning. In *2016 Third European Network Intelligence Conference (ENIC)*, pages 77–84, 2016.

[5] B. Boenninghoff, J. Rupp, R. Nickel, and D. Kolossa. Deep Bayes Factor Scoring for Authorship Verification—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org, Sept. 2020.

[6] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD RECORD*, 30:55–64, 2001.

[7] A. Ellegård. *Who was Junius? A statistical method for determining authorship: the Junius letters 1769–1772*. Gothenburg Studies in English; 13. Acta Universitatis Gothoburgensis, Göteborg, 1962.

[8] O. Halvani, L. Graner, and R. Regev. Cross-Domain Authorship Verification Based on Topic Agnostic Features—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org, Sept. 2020.

[9] F. Johansson, L. Kaati, and A. Shrestha. Detecting multiple aliases in social media. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 1004–1011. ACM, 2013.

[10] F. Johansson, L. Kaati, and A. Shrestha. Time profiles for identifying users in online environments. In *Proc. 1st Joint Intelligence and Security Informatics Conference, IEEE Computer Society*, pages 83–90, 2014.

[11] F. Johansson, L. Kaati, and A. Shrestha. Timeprints for identifying social media users with multiple aliases. *Security Informatics*, 4(1):1–11, 2015.

[12] T. J. Kaczynski. *The Unabomber Manifesto: Industrial Society and Its Future*. Jolly Roger Press, 1995.

[13] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, and B. Stein. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org, Sept. 2020.

[14] A. Q. Morton. *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. Scribner, 1978.

[15] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, E. Shin, and D. Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy (SP)*, pages 300 –314, may 2012.

[16] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. Carvalho, and E. Stamatatos. Authorship attribution for social media forensics. *IEEE Trans. Inf. Forensics Secur.*, 12(1):5–33, 2017.

[18] U. Sapkota, T. Solorio, M. Montes, S. Bethard, and P. Rosso. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics.

[19] R. Schwartz, O. Tsur, A. Rappoport, and M. Koppel. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1880–1891, Seattle, Washington, USA, Oct. 2013. Association for Computational Linguistics.

[20] M. Spitters, F. Klaver, G. Koot, and M. V. Staalduinen. Authorship analysis on dark marketplace forums. *2015 European Intelligence and Security Informatics Conference*, pages 1–8, 2015.

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

[22] O. D. Vel, A. Anderson, M. Corney, and G. Mohay. Multi-topic e-mail authorship attribution forensics. In *Proceedings ACM Conference on Computer Security - Workshop on Data Mining for Security Applications*, pages –. 2001.

[23] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[24] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy Reference Guide, Release 1.5.2, 2020.

[25] J. Weerasinghe and R. Greenstadt. Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification—Notebook for PAN at CLEF 2020. In L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol, editors, *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org, Sept. 2020.

[26] N. Zechner. *A novel approach to text classification*. PhD thesis, Umeå universitet, 2017.