Voting for Authorship Attribution Applied to Dark Web Data

Britta Sennewald
britta.sennewald@unb.ca
University of New Brunswick
Fredericton, Canada

Rainer Herpers
rainer.herpers@h-brs.de
University of Applied Sciences Bonn-Rhein-Sieg
Sankt Augustin, Germany

Marco Hülsmann
marco.huelsmann@h-brs.de
University of Applied Sciences Bonn-Rhein-Sieg
Sankt Augustin, Germany

Kenneth B. Kent
ken@unb.ca
University of New Brunswick
Fredericton, Canada

ABSTRACT

This research is about authorship attribution (AA) within multiple Dark Web forums and the question of whether AA is possible beyond the boundaries of a single forum. AA can become a curse for users that try to protect their anonymity and simultaneously become a blessing for law enforcement groups that try to track users. In this paper, we explore AA within multiple Dark Web forums to determine whether AA is possible beyond the boundaries of a single forum. The analysis revealed that analyzing all features together with a single classifier does not achieve as good results as when they are classified separately and the final result is computed by a voting mechanism. The latter achieves an F1-Score that is up to 44% higher than in the former case. On top of that, the analyses show that the author of a post is at least 94% within the top three most likely candidates. This shows that AA can threaten the anonymity of Dark Web users across the boundaries of different forums.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**; **Natural language processing**;

KEYWORDS

Authorship Attribution, Dark Web, Machine Learning, Natural Language Processing, Voting

ACM Reference Format:
Britta Sennewald, Rainer Herpers, Marco Hülsmann, and Kenneth B. Kent. 2020. Voting for Authorship Attribution Applied to Dark Web Data. In *Proceedings of 30th Annual International Conference on Computer Science and Software Engineering (CASCON'20).* IBM Corp., Riverton, NJ, USA, 10 pages.

1 INTRODUCTION

Authorship attribution (AA) focuses on assigning documents to their corresponding authors. This is very successful when a sufficient amount of text is available. If the length of a text and/or the number of texts per author is small, it becomes increasingly challenging to attribute them correctly to an author [23]. Nonetheless, several researchers e.g., S. R. Pillay et al. [18] or R. Layton et al. [11] focused especially on this typical problem of online texts. It applies to texts within the openly accessible internet (Surface Web) as well as for the Dark Web. The latter is accessible via special technologies such as The Onion Router (TOR), which are becoming increasingly popular [1]. It offers the opportunity to access the internet anonymously, which is interesting for users who want to protect their privacy or circumvent censorship on the internet, but also for those who do not want to be identified when performing illegal activities [21]. Therefore, AA transferred to the Dark Web implies, that posts within Dark Web forums could be assigned to their authors, and thus it means that AA can be a danger to their anonymity.

A good AA algorithm is of interest to law enforcement agencies who, e.g., want to track or find users engaged in illegal activities shown by scientists like M. Yang et al. [25] or M. Sultana et al. [22]. However, the same algorithm can be used to identify users who are dependent on the anonymity of the Dark Web to be able to express their opinions freely and thus avoid the suppression of regimes. Therefore AA can be both a way to track criminals, as well as a danger to privacy on the Dark Web. For these two reasons, it is interesting to investigate how precise posts can be assigned to their authors on the Dark Web, especially between multiple forums.

In this research, four different Dark Web forums are crawled/-scraped to apply AA to posts published by authors that are active in two of these Dark Web forums. For this purpose, different techniques like Natural Language Processing (NLP) and Machine Learning (ML) are used. By doing so, it is possible to determine to what extent posts can be attributed to their correct author, regardless which of the two forums the posts were published within or which username was used. A very important role in this context will also be played by the application of a voting classifier, which will be used in addition to the normal classifiers.

Since usernames within forums can be chosen freely by everybody they are not a reliable way to link user profiles. Hence, within this research, another way that provides more reliability was chosen as a ground truth: Pretty Good Privacy (PGP)- keys that are often used for confidentiality reasons within Dark Web forums and marketplaces. Not all users use the same PGP-key within two or more forums, however, users that do so are linkable by this feature.

CASCON'20, November 10–13, 2020, Toronto, Canada

2 RELATED WORK

Authorship attribution is a difficult task in an environment with hundreds or even thousands of different authors and texts that differ greatly in size. Many researchers already focused on authorship attribution and also on the Dark Web, but only a few researchers have concentrated on a combination of both. Two of them are Ho and Ng, which analyzed stylometric features of texts posted in different Dark Web forums [9]. They tried to connect ten authors within Dark Web forums by extracting stylometric-based features and special *fingerprints* like typical words or typos of an author. Within their analysis, Ho and Ng focused on Support Vector Machine (SVM) classifiers only.

Another research project regarding Dark Web and authorship attribution was undertaken by Spitters et al. [21]. They achieved their results by using a combination of time-based and stylometric features as well as character trigrams. For the classification, they also used Support Vector Machines just as Ho and Ng did [9]. By using all three features together, the actual author was ranked first with a probability of 88% and within the top five most highly ranked candidates with a probability of 97%.

Forum posts also provide another interesting feature for AA, the time when a post was posted. Even if the research of La Morgia et al. [15] is not in the context of authorship attribution, it shows that only the activity of a user can reveal his/her location. The authors analyzed the activity of users of five different dark web forums. In the case of two of them, they already knew where most of the users were coming from. Hence, their goal was to figure out where the users of the other three forums were coming from. To obtain a ground truth regarding how the activity of users appears around the world, they used a Twitter data set with the known origin of all users. After that, they were able to distinguish groups of dark web users from different regions around the globe within a crowd just by analyzing the timestamps of posts. Therefore it can be concluded, that the activity of users that are living in a different timezone is different regarding the time. Hence, this feature might be especially useful when analyzing forums with a global community.

Ashcroft et al. tried to match multiple aliases of the same user within a data set of an Irish web forum and within a Twitter dataset [2]. This research did not focus on the Dark Web environment, but tried to match users within different domains by using AA techniques. The authors used stylometric and time-based features, but they also added so-called emotion-based or Twitter-specific features to their analysis. Within their analysis, the authors used three different machine learning classifiers AdaBoost, Support Vector Machine, and Naive Bayes (NB).

Linking users within different Dark Web forums is one of the main tasks in this research and is quite challenging. Fortunately, although not in the context of authorship attribution, Me, Spagnoletti and Pesticcio focused on the relationship between different TOR marketplace users by concentrating on their PGP-keys [12]. Within Dark Web marketplaces, PGP-Keys are often used to provide confidentiality within transactions. Hence, the public key of vendors and buyers can often be found in their user profile. Even if the original research problem of this paper does not belong to AA, it points out that Dark Web users are using PGP keys for their transactions and can be linked by these keys.

Pillay and Solorio also worked on AA of Web Forum posts [18]. They used stylometry features (lexical and syntactical), statistical language models, clustering, and different machine learning algorithms in their work. Clustering was applied to use the output as meta-features for identifying the authors. The results show, that their approach is able to classify posts of five authors with a probability of around 90% correctly by using BayesNet. However, by an increasing number of authors, the C4.5 algorithm to create a decision tree seems to be the better choice than BayesNet. Besides that, the authors observed that when the number of authors increased, those classifiers that incorporate a cluster identifier worked best.

Swain, Mishra, and Sindhu give an overview of recent approaches to AA techniques [23]. They list multiple research projects focusing on this area, especially the category, language, domain, features, and techniques that have been used. This survey shows that Naive Bayes and Support Vector Machine are the most commonly used classifiers, English the most frequently analyzed language and lexical and syntactic features the most popular features.

3 DATASET

One of the most important parts of this research was to create a suitable dataset for AA analysis. Thus the following section focuses on how this dataset was created, which Dark Web forums were used, and which features were extracted.

3.1 Creation of the Dataset

Developing a crawler for the Dark Web is not as straightforward as for the Surface Web. Therefore, in this section, the most important key aspects for crawling/scraping Dark Web forums that were essential for creating this dataset are briefly described. Some of them are strongly influenced by Gwern Branwen's experiences when creating his Dark Web dataset between 2012 and 2015 [6].

First of all, connecting to a Dark Web website is completely different than connecting to a Surface Website. The TOR network is accessible over the TOR browser or a TOR proxy that can be used by a crawler. To increase the speed of the crawling process multiple TOR proxies were set up to allow multiple TOR sessions in parallel.

However, Dark Web websites tend to protect themselves from being crawled or attacked, e.g., by a Distributed Denial of Service (DDoS) attack [14], more rigorously than websites on the Surface Web. Therefore, the timing of requests and thus the speed of the crawler is an extremely challenging task within the Dark Web. Too many requests per minute might lead to a detection of the crawler, whereas too few requests will result in a slow crawl. Thus, a trade-off between both factors needs to be found, which unfortunately is a trial and error process.

In addition, most Dark Web forums do not allow users to access the content of the forum without being registered. Fortunately, having an account on a Dark Web forum is often linked with the possibility to adjust the settings for the forum's outward appearance. For example, choosing the highest possible number of posts per page will result in a faster crawl as there are fewer pages to crawl. Furthermore, an approach that contains blacklisting as well as whitelisting needs to be set up, e.g., to avoid accidentally being logged out during the crawling process. The last important aspect was to separate automatic and manual processes as all forums

Table 1: Statistics of all four crawled Dark Web forums, showing an overview of the dataset used in this thesis. The number of PGP-Key owners within TMG refers to only those that could be matched.

Forum name	*Number of posts*	*Number of users*	*Number of users with PGP Key*	*Total number of files crawled*
DNMA	75,165	10,489	277	20,645
TH	225,135	26,502	1,900	55,106
TMG	201,538	5,121	193+	15,678
Dread	385,839	63,299	2,943	163,943

require solving a captcha when signing in, or logging in. Since the captchas were not easy to solve automatically, this had to be done manually, whereas the final crawl was completely automatic.

3.2 Dark Web Forums Used

The number of active users in the dark web forums found between October and December 2019 within the context of this research ranged either between a few hundred or between a thousand and more. Since the probability of finding users who are active in two or more forums is expected to be higher when concentrating on those forums that seem to be the most popular, only forums with more than 1000 active users were selected. However, in future work, this threshold could be lowered to also include smaller forums with only a few hundred users to increase the size of the data set [1].

At the end of 2019 there were fewer than 10 Dark Web forums found with a large community (around 1000 active authors or more). Unfortunately, the number of those forums that allow users to publish their PGP keys in their user profiles, was even smaller. In the end, only four forums fulfilled the requirements for this analysis, which are presented in Table 1.

3.2.1 DNM Avengers. This is the smallest Dark Web Forum that was crawled within this research. It focuses mainly on drugs, but there are also some threads about more general topics, politics, security, cryptocurrency or Dark Web marketplaces.

3.2.2 The Majestic Garden (TMG). This onion service is a mixture of forum and marketplace. It has some threads dealing with general topics but the main focus is on drugs as well. TMG has over 40,000 users in total (end of 2019) but only around 5,000 that are active (wrote at least one post) in the forum. Compared to the three other forums, it is the only one where users are not able to access the user profiles of others. Therefore, users post their public PGP keys within multiple threads to share them with others. Unfortunately, it's very time-consuming to check all posted PGP keys manually, whether they are complete or unusable or whether they have been posted twice or even more times. Hence, due to simplicity, only those users that have a PGP key and are linkable to one of the three other forums are manually checked and counted as users that own a key in Table 1. Hence, 193+ indicates that there are more users with a PGP key within this forum, but their exact number was not calculated.

[1]Due to privacy concerns, the data set created for this project is available by request only on GitHub [7].

3.2.3 The Hub (TH). TH is the sister-forum of The Majestic Garden, containing approximately 225,100 posts and 26,500 active users. As its name already indicates, TH is a kind of central point with many threads with discussions about other Dark Web sites, mainly marketplaces. Additionally, it contains many threads regarding security and vendor reviews.

3.2.4 Dread. Dread is the biggest forum crawled within this research with over 63,000 active users. In contrast to the others, it does not have a specific topic. It is more like a platform for everyone that wants to ask questions or talk about various topics. Besides that, the design of this website strongly resembles Reddit, which many people already know from the Surface web. Dread recently faced significant DDoS attacks and thus has extremely strict DDoS protection, which makes it difficult to crawl.

3.3 Features

Features that are extracted from the given data are the basis for an AA analysis. The feature categories used in this research, are influenced by related work or are established based on scientific interest (language model). Another category called social-based features (e.g., the usage of quotations of other user comments) was used in previous work but did not contribute well to the final results. Therefore, this feature category was excluded from this research. However, there might be other features (e.g., transforming text to an image) that are worthy of interest but that are not considered here. These could be analyzed in future work.

The features used in this project are listed in Table 2. More detailed information can also be found in [20]. The four feature-categories used in this research, as well as their corresponding subfeatures, are explained in the following.

3.3.1 Lexical-based Features. Term frequency, also known as Bag of Words (BOW), or Term Frequency Inverse Document Frequency (TF-IDF) features are called lexical-based features in this research. They are often used in the context of AA [18], [11]. The main focus of these features is not on the topic but rather on the frequency of words within a text. However, the TF-IDF approach tries to overcome a typical problem of the BOW approach that rarely used words (that might be the most interesting ones) are shadowed by more frequently used words.

3.3.2 Stylometric-based Features. Stylometric features are frequently used for authorship attribution tasks [21],[2], [9]. The focus of this category is not on *what* an author is writing about, but rather *how* he writes a text. This includes grammar mistakes, typos, emojis, part-of-speech (POS) tags, as well as statistical measurements of an author's writing style, e.g., the number of sentences, words, characters per word, etc.

3.3.3 Time-based Features. This feature category is inspired by La Morgia et al. [15], and Spitters et al. [21]. It contains six subfeatures: the time (hour and minute) when a user is typically active within a Dark Web forum, the date (year, month, and day) that a post was posted, and the day of the week on which a post was written, which might be very important to see whether some users tend to be more active during weekends and others more during weekdays.

Table 2: Features used within this research. For those features that are annotated with a * the sum, mean, median, and standard deviation are computed with regard to every post.

Category	*Feature*	*Extraction Tool*
Lexical	Count Vectoriser	scikit-learn [16]
Lexical	TF-IDF	scikit-learn [16]
Language Model	Word2Vector, GloVe, FastText	Gensim [19]
Language Model	Sentiment (pos./neg./neu./comp)	vaderSentiment [10]
Language Model	LDA and NMF	scikit-learn [16]
Stylometric	11 Emoji categories	RE (Python)
Stylometric	Grammar mistakes	LanguageTool API [8]
Stylometric	Typos	LanguageTool API [8]
Stylometric	35 POS tags	nltk [3]
Stylometric	Number of characters per word*	Python
Stylometric	Number of capital letters*	RE (Python)
Stylometric	Number of small letters*	RE (Python)
Stylometric	Number of punctuation marks*	RE (Python)
Stylometric	Number of abbreviations*	nltk [3]
Stylometric	Number of lowercase-words*	RE (Python)
Stylometric	Number of uppercase-words*	RE (Python)
Stylometric	Number of words with both cases*	RE (Python)
Stylometric	Number of numbers	Python
Stylometric	Number of words	Python
Stylometric	Number of spaces	Python
Stylometric	Number of sentences starting with a capital letter*	RE (Python)
Stylometric	Lexical richness	nltk [3]
Stylometric	Number of Sentences	nltk [3]
Stylometric	Number of Lines	nltk [3]
Stylometric	Number of invisible Characters*	RE (Python)
Time	Minute, hour, day, month, year, day of the week	Python

3.3.4 Language model-based Features. This feature category contains features based on a sentiment analysis, on two topic-modeling algorithms, Latent Dirichlet Allocation (LDA) [4] and Non-Negative Matrix Factorization (NMF), as well as on three language-modeling algorithms, Word2Vec [13], GloVe [17], and FastText [5]. These algorithms are not typically used within AA analyses. However, they were chosen to be part of this research to experiment with different ways to analyze the topic, word embeddings, semantic, and sentiment an author typically uses within a post.

4 METHODOLOGY

AA in the Dark Web is technically hardly different from AA in the Surface Web, but it is very different in terms of the conceptual view and the underlying conditions. In the Surface Web, there are large data sets (e.g., datasets based on Twitter) that can be used for AA. In many cases, they contain the full name of the authors which can be used as ground truth for a supervised ML approach. The Dark Web, on the other hand, contains relatively few data sources with posts from users who want to hide their identity (ground truth). The latter makes it nearly impossible to combine training data from the Surface Web and test data from the Dark Web for a supervised ML approach and therefore to overcome the problem of a limited amount of data in the Dark Web.

The authorship attribution (AA) analysis within this research is based on ML tools provided by scikit-learn version 0.22 [16] and is very extensive, which is why the following sections are very important to understand and interpret the results described in Section 5. Due to the immense computational costs of the analyses, hardware provided by the Platform of Scientific Computing at Bonn-Rhein-Sieg University of Applied Sciences was used for this research [24].

Table 3: Remaining forum combinations and number of corresponding authors after filtering out all authors with less than 50 posts in both forums.

Name	*Forum combination*	*Number of authors*
FC-1	DNMA & TH	2
FC-2	DNMA & TMG	2
FC-3	DNMA & Dread	7
FC-4	TMG & Dread	10
FC-5	TH & Dread	17
FC-6	TMG & TH	20

Table 4: The number of words per author. All values are averaged over all authors within the respective forum combinations.

FC	*Median*	*Mean*	*Standard Deviation*	*Min*	*Max*	*Sum*
FC-1	484	626	734	2	5697	223659
FC-2	121	238	467	4	4732	108605
FC-3	158	366	482	8	3963	111012
FC-4	168	243	390	4	2775	74720
FC-5	132	307	530	2	4016	111028
FC-6	145	268	560	2	4992	131828

4.1 Preparations

Before the analysis could start it was crucial to select suitable authors from the dataset, as well as to decide which preprocessing tools and which classification algorithms should be used. Detailed information about all preprocessing steps and classifiers used within this research can be found in [20].

4.1.1 Selection of Authors. One of the first steps to be able to start the analysis was to find appropriate authors. In this case, the term appropriate refers to authors that can be linked via a PGP-key. On top of that, a trade-off needed to be found between the number of posts that have been written by an author within a forum (the more the better) and the total number of authors that remains for the analysis (the more the better). Therefore in this research, only authors that have written at least 50 posts in two forums were considered to be appropriate. The remaining forum combinations with more than one candidate author, as well as the number of remaining authors, are listed in Table 3. In addition, some text statistics are available in Table 4.

4.1.2 Preprocessing Tools. Before data is fit into a classifier, it is often beneficial to preprocess it first to either fulfill the requirements of a classifier or just to improve the final results. In this research, three different kinds of preprocessing are tested: standardization, normalization or no preprocessing at all. Standardization is used because some estimators are sensitive to the distribution of the data they are fitted with. Normalizing is especially useful when the similarity between a pair of samples should be computed and is

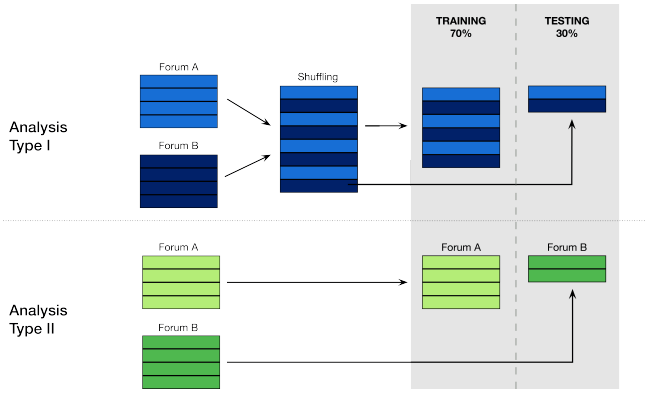

Figure 1: Visualisation of analysis type I and analysis type II

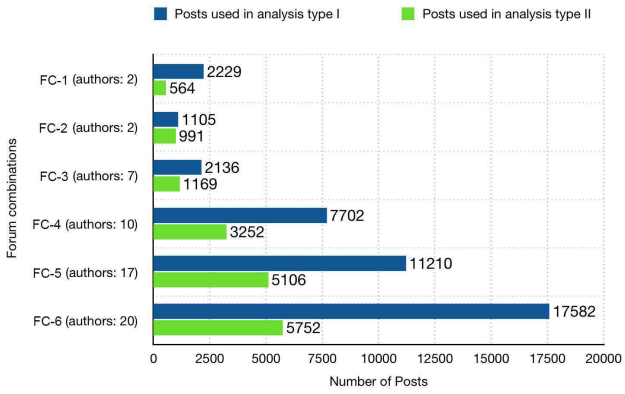

Figure 2: Each bar shows the sum of all posts from all authors within the corresponding forum combination and type of analysis.

often used for text classification, e.g., in the TfidfTransformer from scikit-learn [16].

4.1.3 Classifiers. There are several different classifiers available in scikit-learn. Some of them can be linked to the three classifier categories SVM, Naive Bayes, and Decision Trees, which are in the main focus within this research. However, other classifiers like K-nearest neighbors (KNN) or a Multilayer Perceptron (MLP) are used in this analysis.

4.2 Final Setup of the Analysis

The analysis of this research is extensive because of the intention to analyze the data in as many ways as possible to find the most appropriate one. First of all, each of the six forum combinations listed in Table 3 is analyzed in two different ways (see Section 4.2.1 and 4.2.2). In both types, each forum combination is once analyzed with an unbalanced dataset (with the full amount of data) and once with a balanced dataset (see Section 4.2.3). Furthermore, within each analysis type, each forum combination is analyzed with three versions of the dataset. Each of these three versions contains only those authors that wrote a specific minimum of posts (see Section 4.2.3).

4.2.1 Analysis Type I: Combined Analysis. The most common technique for AA is to extract a text corpus that includes all texts from all authors and split this corpus into a training set and testing set. This is done in the first part (type I) of the analysis by combining each forum pairing as listed in Table 3 into a single dataset. After that, the posts are mixed and split up into 70% training data and 30% testing data. The advantage of this type of analysis is that all posts from all linked authors can be used. However, a major disadvantage is that the proportion of posts from forum A and B vary significantly from author to author. Therefore, the main focus of this analysis is on the feasibility of AA, based on posts from different Dark Web sources. The results of this analysis type will always be visualized in blueish colors in all the figures in this paper.

4.2.2 Analysis Type II: Separate Analysis. The second part of the analysis is based on training sets that contain only posts from forum A of a given forum combination and testing sets that contain only posts from forum B of the same forum combination. The differences between analysis type I and II are visualized in Figure 1. Within analysis type II, the forum that has more posts per author on average is used for the training set and the other for the testing set respectively. Furthermore, the proportion between the two sets remains the same as in type I (70%/30%), which unfortunately leads to a high loss of data. The extent of this problem is illustrated in Figure 2. However, when the number of posts of an author has to be reduced to maintain the ratio, then only the longest posts were chosen for the corresponding data set.

This type of analysis can reveal which features can achieve good results even when the author might have changed some of their typical behaviors between the training and testing forum. Therefore the focus of this analysis is on the suitability of the extracted features. The results of this analysis type will always be visualized in greenish colors in all figures in this paper.

4.2.3 Sub-analyses. Both analyses (type I and II) are further divided into several sub-analyses. In general, the more text from an author that exists, the better is the probability of a successful AA analysis. Thus, there are three different sub-analyses for each analysis type. In the case of analysis type I, this is an analysis with all authors, one with only those authors that wrote more than 500 posts, and the last with only those authors with more than 1000 posts. As the total number of posts per author is lower in analysis type II, one sub-analysis is based on all authors, the second on all authors that wrote more than 200 posts (summed over both forums), and the last on authors that wrote more than 400 posts.

A major problem of both analysis types, as well as the previously mentioned sub-analyses, is that some authors wrote a huge number of posts whereas others just wrote only a few hundred or fewer. Figure 3 visualizes the situation in analysis type I only, but it is similar to that in analysis type II. To analyze the effect of this problem on an AA analysis, all datasets are analyzed twice; one time unbalanced and the other time balanced. Balanced means that the number of posts of all authors is limited to the number of posts written by the author with the fewest posts within the dataset.

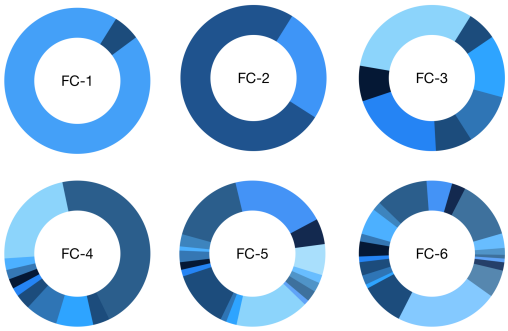

Figure 3: The proportion of the number of posts for each author within the six datasets of analysis type I, where each color represents a different author.

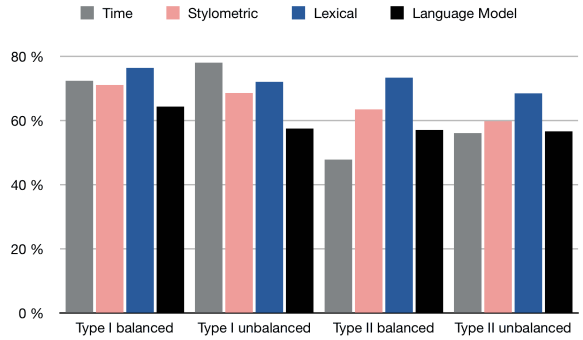

Figure 4: The average F1 scores achieved by the four feature categories calculated over the results obtained from the analyses of all six forum combinations.

Similar to the procedure in analysis type II, only the largest posts are kept when shrinking the number of posts of an author. In the case of analysis type I, also the proportion of posts from forum A and B was balanced as much as possible to focus on the main research question, which was AA of posts written in two different Dark Web forums.

5 RESULTS

The results described in this section are selected results for each forum combination and each feature category within each analysis mentioned in Section 4. In this research, the best results of an analysis and/or classifier are considered to be those with the highest F1-score and the highest accuracy. This is based on the fact that a high recall, as well as a high precision, are essential for AA in the Dark Web to assign a post to its correct author as reliably as possible. Since the F1 score is a kind of average between precision and recall, only this score is chosen as an evaluation criterion.

When focusing on Figures 5 and 6 it can be observed, that AA within this analysis becomes more difficult with more candidate authors. However, there is a significant difference between the results of analysis type I and analysis type II, which is most significant for datasets with more than two authors (FC-3 to FC-6). The results of analysis type II (Figure 6) are, in general, 10% to 20% worse than those achieved within the analysis type I (Figure 5). This tendency can also be found in the remaining analyses with a focus on authors that have written comparatively many posts.

The results achieved by the different feature-categories differ significantly among each other and between the different types of analyses. Therefore, in the following subsections strengths and weaknesses of the different categories that can be concluded from the given results of FC-5 and FC-6 are described. These two forum combinations are chosen for this more detailed analysis because they represent the challenges and difficulties for a successful AA analysis. However, a brief overview of the *average* results (F1-score) of the feature categories within the analyses of *all* forum combinations is shown in Figure 4.

5.1 Time-based Features

Time-based features belong to the most accurate features of all within both analysis types of FC-1 and FC-2 but unfortunately not for the other four forum combinations with a higher number of authors (see Figures 5 and 6). However, a detailed analysis of the results revealed that the correct author of a post is determined 88% (FC-6) and 91% (FC-5) within the three most likely candidates when considering time-based features only and all authors within type I. Except that, in 40 out of all 52 analyses, the time-based features were classified best by tree-based classifiers like the ExtraTrees classifier and RandomForest classifier.

Within FC-5, one main reason for good performance of the time-based features is the number of posts per author. However, the results of FC-6 show that it does not necessarily mean that an author is unidentifiable when they have written only a few posts. This indicates two facts: in general, the more candidate authors there are, the higher the number of posts per author is needed to identify an author using time-based features. On the other side, some authors do not need a high number of posts because their daily rhythm seems to be so atypical that they stand out easily. When taking a look at the balanced analysis that considers the longest posts of all authors, the F1-score of those authors that wrote many posts drops significantly. This is because the number of posts is reduced to a minimum number of posts written by an author within the dataset. On top of that, when balancing the dataset by choosing posts at random, the same tendency occurs. This leads to the conclusion that there is, in general, no connection between the length of a post and the time when a post is written. As visible in Figure 6 the results achieved by analysis type II are significantly lower than within analysis type I. The most obvious reason for the poor results within type II would be that there are simply not enough posts for each author to be able to find all daily rhythms. As the results within analysis type I are comparatively high, it seems that authors tend (at least) to be active to a similar time within their favorite forum in that they have written the most posts.

5.2 Stylometric-based Features

Compared to the time-based features, the results of the stylometric-based features are worse within analysis type I whereas they are,

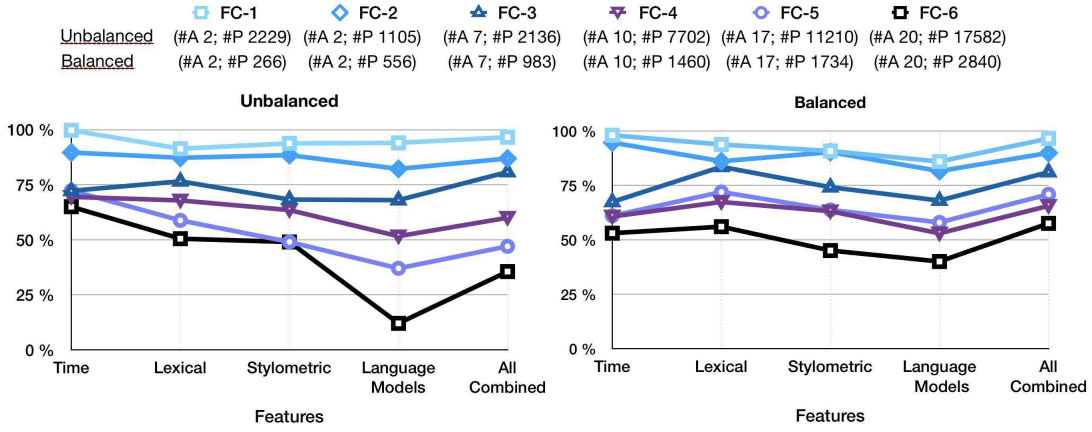

Figure 5: Overview of the F1-Score achieved by analyzing the complete datasets of all forum combinations for each feature category with analysis type I by using either an unbalanced dataset or a balanced dataset. #A denotes the number of authors within a forum combination and #P the number of posts.

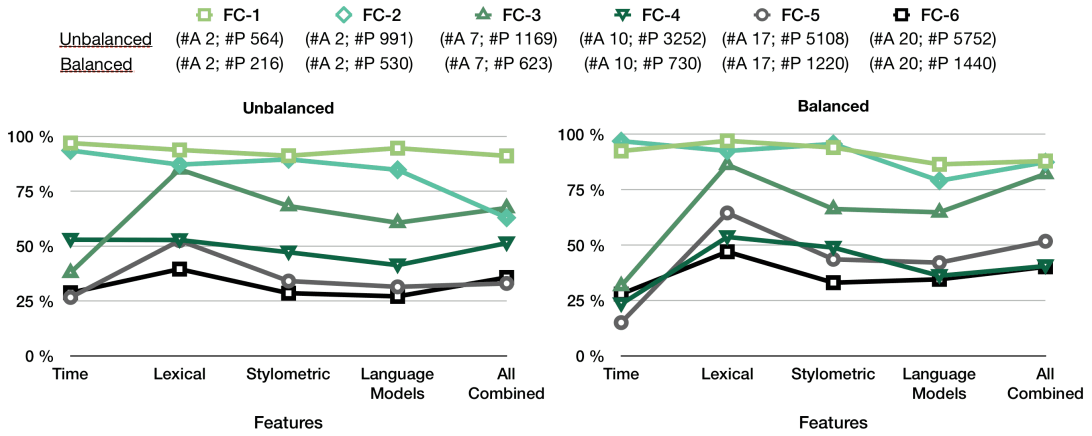

Figure 6: Overview of the F1-Score achieved by analyzing the complete datasets of all forum combinations for each feature category with analysis type II by using either an unbalanced dataset or a balanced dataset.

in general, better within analysis type II (see Figures 5 and 6). In addition to the results shown in these two charts, the probability that the correct author can be found within the top 3 most likely candidates within FC-5 and FC-6 is around 50% for type II and ranges between 70% and 83% for type I. However, the classifiers that achieved the highest results by analyzing this feature category, were the LinearSVC (26/52), the ExtraTreesClassifier (11/52), and the MLP Classifier (9/52).

One of the most important factors for a high F1-score when analyzing stylometric-based features seems to be the number of posts. This is based on the observation that the F1-score of authors that could be identified best is more negatively affected when the dataset is balanced (no matter if only the longest posts are chosen or not). Therefore, the stylistic pattern that makes these authors expose most, can only be found when analyzing a great number of posts and thus a mixture of long and short posts. The remaining analyses revealed, that the more posts per author on average are included in the dataset, the more effective it is to focus on the writing style in longer posts. Furthermore, the results achieved by poorly detectable authors show that balancing the dataset and focusing on only the longest posts leads to an improvement of their score up to 44%. Therefore, a tradeoff needs to be found that keeps as many posts as possible for each author from the well-identifiable authors within the dataset, but at the same time, reduce the number of posts as much as possible so that the others also have a chance to stand out.

However, results of analysis type II are around 20% lower than those within analysis type I. The main reason for that seems to be the comparatively low number of posts within the dataset as well as a tendency, to write posts of a different length within the two forums. Therefore, the overall conclusion when considering the results of analysis type I and II is that authors seem to write more passionately in either one or the other forum, which results either in a different style of their posts or in a different number

of posts per forum that are available for the analysis. Both cases are a problem for the stylometric-based features especially within analysis type II.

5.3 Lexical-based Features

The results shown in Figures 5 and 6 prove that the lexical-based features belong to the best features within this analysis. When an unbalanced dataset is used, then the probability that the correct author of a post is within the top three candidates is at least very close to or above 60% in nearly all cases (analysis type I and type II). On top of that, when the analysis is based on a balanced dataset, the probability of finding the correct author within the top three rises to at least 66% and up to 86%. When taking a closer look at the best classifiers, then it becomes apparent that Naive Bayes classifiers seem by far the most suitable type when analyzing the lexical-based features (in 37 of 52 analyses).

When focusing on the results of each author within the balanced and unbalanced analyses (those that consider all authors) it becomes obvious that an author that wrote long posts stands out from the crowd more easily when considering only a few posts than an author that usually writes many short posts. This means that the length of the posts is the most important influencing factor for the lexical-based features. Furthermore, when comparing the results achieved by an analysis of all authors to those achieved by an analysis that considers only authors with 500 or 1000 posts, an interesting fact can be observed. Focusing on a few longer posts per author by balancing the dataset is, in general, more or at least equally effective than focusing on only those authors that wrote comparatively many posts.

In general, the lexical-based features are suited comparatively well for authorship attribution within analysis type II. As the overall score of the lexical-based features within analysis type II is better than that of the stylometric-based features, it seems like authors rather tend to write about the same things or at least use similar words within two different forums than to use the same writing style when posting a post.

5.4 Language Model-based Features

The language model-based features, combined together into a single dataset, are not suitable for AA within the Dark Web (see Figures 5 and 6). Datasets with a large number of authors (FC-6) seem to be especially problematic; in the best case the correct author of only 62% of all posts is listed within the top 3 most likely candidates. FC-5 has three fewer authors, which seems to lead to a slightly better probability of finding the correct author within the top 3 candidates (up to 75%). When examining at the classifiers, the LinearSVC (25/52) and the PassiveAgressiveClassifier(13/52) seem to achieve the best results of all classifiers when analyzing the language model-based features only.

There are a few authors that can be recognized better when focusing on language model-based features only but there is no single common reason for all authors. Some have many posts and others do not, the same applies to the length of the posts, some have long posts, others not. However, two small tendencies can be observed. First, it is more likely that an author that has written 500 or more posts can be identified comparatively well. Second, when an author tends to write small posts, then it is more likely that a classifier cannot classify them by using only this feature category.

5.5 Voting

When putting all features together and analyzing them combined, one might expect that the results must increase because all information is now merged. However, real-life experience shows, that this is not the case within most of the analyses in this research. The probability that a post can be correctly attributed to its author when focusing on all features combined within the unbalanced datasets of FC-5 and FC-6 including all authors is comparatively low (40% or less). However, there is a clear tendency that this probability increases when balancing the dataset. In the best case, it rises within analysis type I to 80% and even higher (to 86%) when the top 3 most highly ranked candidates are included.

The question is, where did the potential of the individual features get lost in the joint analysis? The answer is surprisingly simple: each feature category can be classified best by different classifiers. Thus, when putting all features together and classifying them with only one classifier, the results decrease. Therefore, another approach to analyzing all features combined was tested. In contrast to the previous one, the features are not put together in one single dataset. Instead, they are classified as stand-alone by the same classifier with the same classifier parameters. However, this time, the result of each classifier is fed into a final voting classifier. Thus, this final classifier receives four results from four different classifiers for each sample that is used for testing. Out of these results, the voting classifier computes the most likely candidate author for each post. Since it is known from the previous analyses when each feature category works well, weights can be added to the computation so that those classifiers that are more reliable in a given situation than others, have more influence on the final result. E.g., when focusing on analysis type I with an unbalanced dataset, it is known that the time-based features work very well, the stylometric-based and lexical-based features are not as suitable as the time-based features, but still work well, whereas the language model-based features achieved the worst results. This knowledge produces a tendency for which features should be weighted more or less. However, the final weights still need to be determined by experimentation.

Mathematically, the computations made by the voting classifier can be described as in Equations (1) and (2) where $\mathbf{B} \in \mathbb{R}^{n \times j \times k}$ denotes a three-dimensional matrix that contains the probability estimates of all classifiers (the probability of the posts for each author in each model) and $\mathbf{A} \in \mathbb{R}^{j \times k}$ denotes a matrix that contains the weighted averages $a_{m,l}$ of the probability estimates.

$$a_{m,l} = \frac{\sum_{i=0}^{n} b_{i,m,l} \cdot w_i}{\sum_{i=0}^{n} w_i} \; \forall \; a_{m,l} \in [0,1] \tag{1}$$

for all $m \in \{0, ..., j\}$, where j denotes the total number of samples, for $l \in \{0, ..., k\}$, where k denotes the total number of authors, for $b_{i,m,l} \in \mathbf{B}_{n \times j \times k}$, where n denotes the total number of classifiers, and where the vector w contains the weights for each classifier. The final output of the voting classifier can be mathematically described as shown in (2):

$$v_m = \arg \max_{l \in \{0,\ldots,k\}} a_{m,l} \tag{2}$$

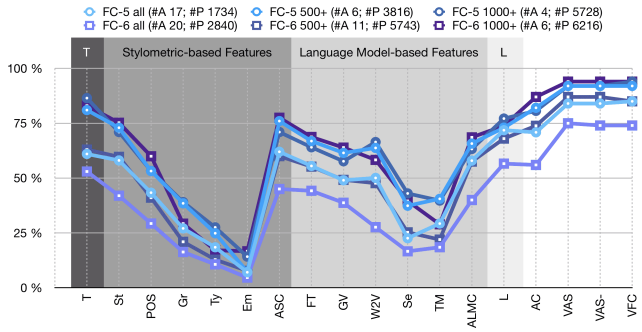

Figure 7: Detailed overview of all results achieved by analysis type I of FC-5 and FC-6 (balanced). #A denotes the total number of authors and #P denotes the total number of posts. Acronyms are explained in Table 5.

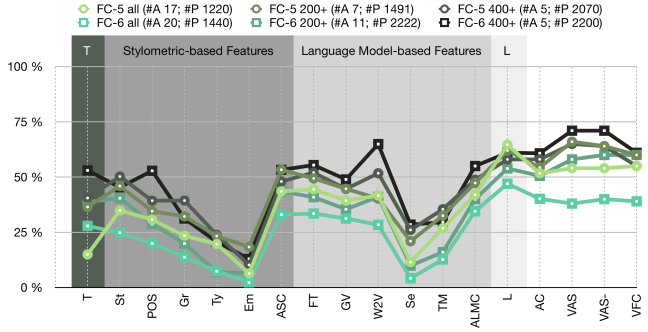

Figure 8: Detailed overview of all results achieved by analysis type II of FC-5 and FC-6 (balanced). #A denotes the total number of authors and #P denotes the total number of posts. Acronyms are explained in Table 5.

Table 5: Acronyms used within Fig. 7 and Fig. 8.

ASC: All Stylometric-based Features Combined	**Em:** Emojis
ALMC: All Language model Features Combined	**FT:** FastText
AC: All Features Combined	**GV:** GloVe
L: Lexical-based Features	**Gr:** Grammar
T: Time-based Features	**POS:** POS-tags
TM: Topic Modeling	**Se:** Sentiment
VAS: Voting all Features Separately	**Ty:** Typos
VAS-: Voting all Features Separately - Emojis etc.	**St:** Style
VFC: Voting Combined Feature Categories	**W2V:** Word2Vec

where *argmax* computes the column index of the author/column with the highest probability and thus the most likely author within each row (sample) of **A**.

Figure 9 shows a significant increase of the results achieved by voting all feature categories instead of putting them all together into one single dataset and analyzing them by a single classifier (Figures 5 and 6). Especially within analysis type I, the results of the previously unsuitable unbalanced datasets increases to 80% in the case of FC-6 which is the largest forum combination with 20 authors. A similar trend can be seen for FC-5, which achieves an F1-Score of 87% when voting the results of all feature categories. Interestingly, the results of analysis type II, as well as those of the balanced analyses, do not increase that much. There might be several reasons for that. All datasets of analysis type II as well as the balanced datasets of analysis type I contain significantly fewer posts. This leads to an increase of the F1-score of the lexical-based features but to a decrease of the score for the time-based features. As the latter ones had the most influence on the voting results of the unbalanced datasets, it is not surprising that the results of the balanced datasets do not increase in the same way. In the case of analysis type II, there were only a few feature categories that passed the 50% limit at all. Thus, the voting classifier is not able to improve the results of this final analysis when most of the feature categories are not able to pass the 50% limit stand alone.

In Figures 7 and 8 a more detailed overview of the results achieved by all features and different types of voting analyses is shown. Voting all Features Separately (**VAS**) denotes a voting analysis of all subfeatures without the results of the corresponding feature category where all subfeatures are combined and analyzed with a single classifier. In Voting all Features Separately without Emojis etc. (**VAS-**) only those subfeatures with the best results are voted. Therefore this analysis does not include the results of the analyses of the emoji, sentiment, topic modeling, grammar, and typo features. As the weights for these features were already quite small within VAS it is not surprising that the voting results of an analysis without them, do not change significantly. The results that are labelled Voting Combined Feature Categories (VFC) are those that were discussed in the previous paragraph. Thus in this case, only four results (one for each feature category) are voted. This version of voting seems to work best for large datasets. However, when the datasets are small, like in the balanced analyses, there is either no difference with the other two voting methods or the results are a little worse. Besides that, Figure 7 shows that in most cases subfeatures that are analyzed separately do not achieve better results than when analyzed combined.

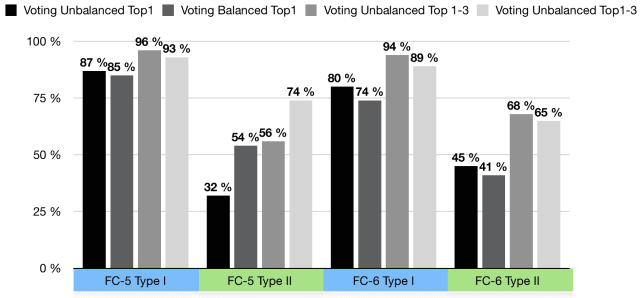

Figure 9: Comparison of the probability that a post is classified to its correct author (top 1) or that the correct author is within the three most likely candidates (top 1-3) when considering all authors and vote the classification results of all features.

6 CONCLUSIONS

The focus of this research was on authorship attribution within multiple Dark Web forums. That AA is feasible within one forum was already shown by e.g., M. Spitters et al. [21]. Thus the main research question was whether it would also be possible beyond the boundaries of a single forum. Therefore, a crawler/scraper was developed that can extract posts from four different Dark Web forums: DNM Avengers, The Hub, The Majestic Garden, and Dread. By comparing public PGP-keys, users were linked between these four forums. The results show that it is (in general) a good idea to reduce the number of posts from authors that wrote significantly more posts than the average. By this, it is more likely to better classify authors with only a few posts. However, even under the challenging conditions with posts originating from different sources, there is still a probability of at least 94% that the correct author of a post can be found within the three most likely authors. This result is achieved by voting the results of four different classifiers that classify four different feature categories. However, it shows that AA is indeed a danger to the anonymity of Dark Web users across the boundaries of different forums. Therefore, all users that want to avoid getting linked via AA should keep the following aspects in mind: a post can most likely be assigned to its author if they tend to have an abnormal or very typical daily rhythm that is reflected in their online behavior, if they tend to write many long or short posts with the same structure, and also even when they write only a few but comparatively long posts with a large number of similar words.

6.1 Future Work

Several aspects could not be realized within this research and should be optimized or extended in future work. Unfortunately, only a few authors could be linked at all between these forums and those that could be linked rarely wrote more than 50 posts in both forums. Thus, it would be desirable to find more popular Dark Web forums that can be used as an additional data source for further validating the research presented. It would also be interesting to analyze whether the improvment in the results within the analyses with a focus on authors with a greater number of posts is rather related to the comparatively high number of posts or to the reduced number of authors. Apart from this aspect, the analysis of ensembling methods like stacking, boosting, or bagging would also bear fruitful results.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support of both the Natural Sciences and Engineering Research Council (NSERC) as well as the New Brunswick Innovation Foundation (NBIF) for their support of the research. Thanks also to Stephen MacKay for editing grammar and style and Andreas Priesnitz for the great support on the University of Applied Sciences Bonn-Rhein-Sieg (BRSU) side to help this research move forward. The authors would also like to thank the Plattform of Scientific Computing at BRSU for providing the hardware needed for the analyses.

REFERENCES

[1] Johanna Amann and Robin Sommer. 2016. Exploring Tor's Activity Through Long-Term Passive TLS Traffic Measurement. In *Passive and Active Measurement*, Thomas Karagiannis and Xenofontas Dimitropoulos (Eds.). Springer International Publishing, Cham, 3–15.

[2] M. Ashcroft, F. Johansson, L. Kaati, and A. Shrestha. 2016. Multi-domain Alias Matching Using Machine Learning. In *2016 Third European Network Intelligence Conference (ENIC)*. 77–84.

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3, null (March 2003), 993–1022.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *CoRR* abs/1607.04606 (2016). http://arxiv.org/abs/1607.04606

[6] Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Sohhlz, Delyan Kratunov, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark Net Market archives, 2011-2015. www.gwern.net/DNM-archives. Accessed: 22-05-2019.

[7] CAS-Atlantic. [n. d.]. Dark Web forum dataset 2019 (DWF-CAS-IVC-2019). https://github.com/CAS-Atlantic/Dark-Web-forum-dataset-2019-DWF-CAS-IVC-2019. Accessed: 21-08-2020.

[8] LanguageTooler GmbH. [n. d.]. LanguageTool - Proofreading Software. https://languagetool.org. Accessed: 09-08-2020.

[9] Thanh Nghia Ho and Wee Keong Ng. 2016. Application of Stylometry to DarkWeb Forum User Identification. In *Information and Communications Security*, Kwok-Yan Lam, Chi-Hung Chi, and Sihan Qing (Eds.). Springer International Publishing, Cham, 173–183.

[10] C.J. Hutto and E.E. Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*. Ann Arbor, MI.

[11] R. Layton, P. Watters, and R. Dazeley. 2010. Authorship Attribution for Twitter in 140 Characters or Less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*. 1–8.

[12] G. Me, L. Pesticcio, and P. Spagnoletti. 2017. Discovering Hidden Relations Between Tor Marketplaces Users. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*. 494–501.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. 1–12. https://arxiv.org/abs/1301.3781

[14] Jelena Mirkovic and Peter Reiher. 2004. A Taxonomy of DDoS Attack and DDoS Defense Mechanisms. *SIGCOMM Comput. Commun. Rev.* 34, 2 (apr 2004), 39–53.

[15] M. La Morgia, A. Mei, S. Raponi, and J. Stefa. 2018. Time-Zone Geolocation of Crowds in the Dark Web. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. 445–455.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, null (Nov. 2011), 2825–2830.

[17] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[18] S. R. Pillay and T. Solorio. 2010. Authorship attribution of web forum posts. In *2010 eCrime Researchers Summit*. 1–7.

[19] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[20] Britta Sennewald. 2020. *Authorship Attribution in the Dark Web*. Master's thesis. University of New Brunswick, Fredericton, NB, Canada.

[21] M. Spitters, F. Klaver, G. Koot, and M. v. Staalduinen. 2015. Authorship Analysis on Dark Marketplace Forums. In *2015 European Intelligence and Security Informatics Conference*. 1–8.

[22] M. Sultana, P. Polash, and M. Gavrilova. 2017. Authorship recognition of tweets: A comparison between social behavior and linguistic profiles. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 471–476.

[23] S. Swain, G. Mishra, and C. Sindhu. 2017. Recent approaches on authorship attribution techniques — An overview. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Vol. 1. 557–566.

[24] Bonn-Rhein-Sieg University. [n. d.]. Platform for Scientific Computing at Bonn-Rhein-Sieg University. https://wr0.wr.inf.h-brs.de. Accessed: 08-05-2020.

[25] Min Yang and Kam-Pui Chow. 2014. Authorship Attribution for Forensic Investigation with Thousands of Authors. In *ICT Systems Security and Privacy Protection*, Nora Cuppens-Boulahia, Frédéric Cuppens, Sushil Jajodia, Anas Abou El Kalam, and Thierry Sans (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 339–350.