Natural Language Engineering **21** (4): 497–518. © Cambridge University Press 2013
doi:10.1017/S1351324913000272

Authorship analysis of aliases: Does topic influence accuracy?

ROBERT LAYTON[1], PAUL A. WATTERS[1] and RICHARD DAZELEY[2]

[1]*Internet Commerce Security Laboratory, University of Ballarat, Australia*
e-mails: r.layton@icsl.com.au,p.watters@ballarat.edu.au
[2]*Data Mining and Informatics Research Group,*
University of Ballarat, Australia
e-mail: r.dazeley@ballarat.edu.au

(*Received 19 July 2012; revised 2 September 2013; accepted 7 September 2013; first published online 8 October 2013*)

Abstract

Aliases play an important role in online environments by facilitating anonymity, but also can be used to hide the identity of cybercriminals. Previous studies have investigated this alias matching problem in an attempt to identify whether two aliases are shared by an author, which can assist with identifying users. Those studies create their training data by randomly splitting the documents associated with an alias into two sub-aliases. Models have been built that can regularly achieve over 90% accuracy for recovering the linkage between these 'random sub-aliases'. In this paper, random sub-alias generation is shown to enable these high accuracies, and thus does not adequately model the real-world problem. In contrast, creating sub-aliases using topic-based splitting drastically reduces the accuracy of all authorship methods tested. We then present a methodology that can be performed on non-topic controlled datasets, to produce topic-based sub-aliases that are more difficult to match. Finally, we present an experimental comparison between many authorship methods to see which methods better match aliases under these conditions, finding that local *n*-gram methods perform better than others.

1 Introduction

Social media commonly uses the notion of *aliases* to preserve the privacy of users, creating a separation between their 'real world' and 'online' personae. An 'alias' is a moniker that is (often self-) assigned to a user, identifying that user without revealing their true, 'real world' identity. A commonly used alias is for e-mail accounts; while some corporate e-mail address policies specify the incorporation of a real name, such as `firstname.lastname@company.com`, many 'web mail' providers allow users to select their own e-mail address, resulting in aliases such as `hunter2@email.com`.

With the rise of social media, the use of aliases has increased dramatically. While some sites (such as Facebook) request that real names be used, there is typically little or no verification of such identities during registration. As a result, one person can register for these services under different names or different aliases.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

While the use of aliases can have legitimate purposes, some user classes (including trolls and cyberbullies) use them to evade responsibility for their online activities (Stabek, Watters and Layton 2010). Aliasing can defeat some access controls, such as banning malicious users, since they can simply re-register under a new alias and continue posting. By removing the link between actual and online identity, civil and criminal investigations can be impeded, because the true author of a defamatory post, for instance, may not be readily identified from their alias.

Thus, there is a need (in some circumstances) for reliable alias matching to be performed. Being able to match aliases and ultimately link them to a known identity would enable attribution, and would likely reduce the incidence of some forms of cybercrime that rely upon anonymisation (Layton and Watters, 2009). This approach is consistent with the situational crime prevention principle of increasing the risk by reducing anonymity (Clarke, 1997).

This desire to enhance law enforcement capability needs to be balanced against legitimate desires for privacy and anonymity, allowing citizens to speak freely on the Internet against tyranny or oppression (Ureche, Layton and Watters 2012), particularly in repressive states or those which restrict Internet usage. As noted in Narayanan *et al.* (2012, p. 1), 'a right to anonymity is meaningless if an anonymous author's identity can be unmasked by adversaries'.

Network-based methods also exist that are capable of identifying a user responsible for a posting on the Internet – which we refer to as direct methods – by tracing the source of the network packets to its origin. Countermeasures, such as The Onion Router (TOR), can be readily used to mask the provenance of communication on the Internet. For the purposes of this research, we assume that such *direct* methods of linking aliases are generally infeasible, and therefore, *indirect* methods are required (Watters *et al.* 2012). This assumption is a realistic one, due to the continuous improvement of both sides of attribution and anonymisation technology – engaged, in a sense, in a war of attrition.

In this paper, we refer to the original creator of a document, regardless of the alias used, as the author. The pseudonym that the document is noted as being created by is the alias. When we split an author's documents into groups under the guise of being by two aliases, we refer to these as sub-aliases, to emphasise that these are created within the methodology through differentiation, and not an actual alias chosen by the author.

1.1 Research objective

Research into the alias matching problem often requires generating data, as real-world datasets containing authors attempting to hide behind multiple aliases are difficult to find and/or reliably generate. Previous research has generated multiple sub-aliases for an author by taking all documents under a single alias, and randomly relabelling them. We refer to this process as generating 'random sub-aliases'. Research using this approach has often generated very impressive results, typically with accuracy levels above 90% (Novak, Raghavan and Tomkins 2004) and exceeding 60% in difficult scenarios (Layton, McCombie and Watters 2012).

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

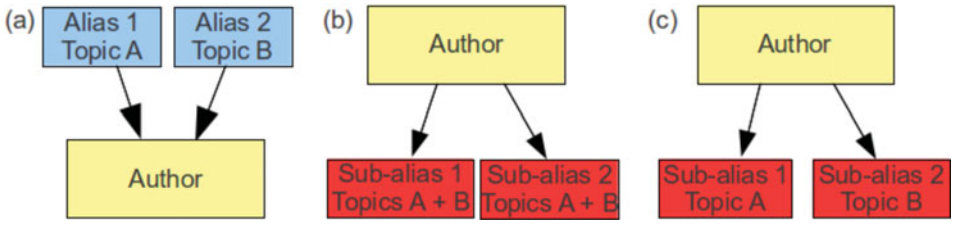


Fig. 1. (Colour online) Comparison between (a) the real-world problem, (b) the idealised problem using random sub-aliases, and (c) the idealised problem using topic-based aliases.

Unfortunately, this solution does not sufficiently model the real-world problem. If an author is hiding behind multiple aliases, it is likely that they are discussing different topics, deliberately choosing which alias to use (and not choosing randomly which alias to use). This distinction is shown in Figure 1. This means that sub-aliases are likely to be very different to each other with regard to topic.

In this paper, we investigate whether this is a major problem in terms of accuracy and applicability, and whether topic-based sub-aliases differ significantly from random sub-alias generation.

1.2 Contributions

In research on alias matching, we aim to create techniques to identify when postings made under multiple aliases are written by the same author. There has been previous research on this application, however, we show that previously used methods drastically reduce in accuracy when sub-aliases are created based on topic (and not randomly). We believe that this is a major limitation to the applicability of existing techniques to real-world problems, and thus, we present a methodology for generating sub-aliases by topic, even if the topics are not known for the datasets. The proposed methodology uses a word-based document clustering technique to split the documents into two, approximating multiple topics. We also present a semantically aware version of the methodology for validation. We finally perform an experimental comparison using a number of datasets from online social media.

This paper makes three major contributions. Firstly, we support our assertion that random sub-aliases generation increases performance over topic-based sub-aliases, by using a topic controlled dataset. This experiment shows that randomly generated sub-aliases are much easier to classify than topic-based sub-aliases. In a real-world application, we could expect significant variance by topic, which should be replicated in experiments in this area. Secondly, we show that generating topic-based aliases using a topic clustering methodology is effective at generating difficult to match aliases, modelling the use case of topic controlled datasets, even when the original dataset does not contain topic information. This allows for the more thorough testing of methods for this application, bridging the gap between the idealised and application scenarios. Thirdly, we test a large number of authorship models using this methodology and show that local *n*-gram (LNG) models perform better than feature-based approaches for this type of application. The reduction in performance

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

was less significant for LNG methods and the overall accuracy was much higher, suggesting that LNG methods are more robust than other types tested.

2 Literature review

Alias matching is a relatively new field of research, but draws heavily upon methods from authorship analysis. Novak *et al.* first investigated the problem of identifying when multiple online aliases belonged to the same author using authorship analysis of the text within their messages (Novak *et al.*, 2004). They used several types of features, including word distributions, misspellings, punctuation characters, emoticons and function words. Using these features to form vector space representations of the documents written under each alias, the Kullback–Leibler (KL) divergence was then used to determine how similar two aliases were. For their experiments, one hundred aliases were used, after which the documents were split into two sub-aliases. The testing corpora were collected from an online forum, composed of approximately 3,000 words per alias. Each sub-alias was then compared against the other 199 sub-aliases. The most similar sub-alias was predicted as the match. This methodology was able to achieve over 90% accuracy in some of the experiments.

There has been other research on linking aliases using methods other than authorship analysis. Anti-aliasing was performed by analysing e-mail address usage by authors across different websites (Holzer, Malin and Sweeney, 2005). Overall, these methods of linking aliases are highly relevant to the practical problem. In this research, we consider these techniques to be complimentary to an investigator interested in matching aliases. However, the experiments in this research will assume that the only information that can be used is the posts themselves, and confine the research to using authorship analysis methods.

Other research, such as that by Schein *et al.* (2010), has investigated the use of topic to inform authorship attribution accuracy. The separated topics when performing cross-validation, allow for the evaluation of authorship methods to be compared based on unseen topics. This approach has some overlap with the goals of this research; however, we focus on the problem of alias matching instead of authorship attribution, and specifically address the instances when topic information is not available.

2.1 Authorship analysis

The problem of anti-aliasing using authorship analysis can be seen as an abstraction of the problem of *authorship attribution*. Authorship attribution is the supervised learning problem of attributing a single document of unknown authorship to one of a set of candidate authors. There is a wide range of research into this problem, and the techniques can be separated into two major categories: vector space models and LNGs.

Vector space models use an ordered set of features F of size d to create a representation $x \in \mathbb{R}^d$ of each document in the testing corpora. This representation

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

can then be used as input to a supervised classification algorithm such as Support Vector Machines (SVM) or used with a distance metric to determine the distance between different documents (under the assumption that a lower distance gives a higher likelihood of co-authorship). The initial choice of features is determined by the investigator and includes features such as the frequency of certain function words, the mean length of sentences or the occurrence of a signature in an e-mail.[1] There are a significant number of features, with Rudman now famously declaring in 1998 that nearly 1,000 features had been used in the authorship analysis literature (Rudman, 1998). Since then, many more features have been used. One of the most thorough descriptions of the most commonly used features is given in Zheng *et al.* (2005). The previously mentioned research of Novak *et al.* (2004) used a vector space model for representing the documents, as have many other papers in the field (Solorio *et al.*, 2011).

LNG models instead aim to determine an author profile for writing habits, using the distribution of character *n*-grams in the author's messages. For a document of unknown authorship, a similar *document profile* is created, and then the profiles are compared. The first technique of this type utilised was the Common *n*-grams (CNG) methodology (Kešelj *et al.*, 2003), which profiled an author or document using the top L most frequently occurring *n*-grams (L is usually in the thousands while n is usually between two and five inclusive). The distance between two profiles, P_1 and P_2, is given in (1).

$$K(P_1, P_2) = \sum_{x \in X_{P_1} \cup X_{P_2}} \left(\frac{2 \cdot (P_1(x) - P_2(x))}{P_1(x) + P_2(x)} \right)^2 \tag{1}$$

Since then, several variations on this scheme have been released. The Source Code Author Profile (SCAP) methodology used the same profile method as CNG, but the similarity between profiles was computed as the size of the intersection (i.e. *n*-grams appearing in both profiles) (Frantzeskou *et al.*, 2007; Layton, Watters and Dazeley, 2011b). Recentred Local Profiles (RLP) used a language default profile to recentre the values within each profile, and used the top L most distinctive *n*-grams, where distinctive was the absolute distance between the frequency of the *n*-gram in normal usage, compared to that author or document (Layton, Watters and Dazeley 2011c). Stamatatos' $d1$ and $d2$ measures are improvements designed to work with unbalanced datasets (Stamatatos, 2007). CNG-WPI (Weighted Profile Intersection) is an improvement to SCAP which weights the *n*-grams based on the number of documents they appear in (inferring the likelihood of the *n*-gram to appear in both profiles) (Escalante, Montes-y Gómez and Solorio, 2011). LNG methods have also been used for unsupervised methodologies, outperforming feature-based models (Layton, Watters and Dazeley 2011a).

Luyckx and Daelemans (2011) investigated the impact of several components of the authorship attribution problem. The effect of the number of candidate authors and the amount of training data available was tested, finding that more data and

[1] The occurrence is usually used, but the text of the signature is typically removed to ensure that it does not make the classification easier than it would otherwise be.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

fewer candidate authors give better results. While unsurprising at a high level, this was the first time such effects have been quantified, and also highlighted the fact that topic variation in texts appears to play a substantial role. While not solidly determined, this *topic effect* was discovered by examining whether a dataset with fewer topics and less data often had higher accuracies than a dataset with more topics and more data.

2.2 Topic modelling

In this section, we will briefly overview methods for modelling the topic of documents. As this research will investigate the impact of creating sub-aliases based on topic difference, it is necessary to understand how topic is modelled. The aim is not to be exhaustive; for in-depth coverage, we recommend Sebastiani's (2002) review, and the more recent work by Aggarwal and Zhai (2012).

Topic modelling[2] is the task of creating a representation of documents to build a classifier or clusterer based on their topic or category. While recent advances have improved the accuracy of such systems, many older methods still provide sufficient accuracy for many applications. A prime example of this is the bag-of-words model. In the bag-of-words model, a matrix X is created such that $X_{i,j}$ is the frequency of word i in document j. In this method, the position, relationship and interaction between words is ignored. While simplistic, this model has been shown to perform quite well, and is often used as a baseline for testing other approaches (Aggarwal and Zhai, 2012).

One method for improving the weighting of a bag-of-words model is to use the *tf-idf* weighting scheme. In *tf-idf*, the weight of a frequency is proportional not only to the number of times it appears in a given document, but also the number of documents in the corpus it appears in Salton and Buckley (1988). For a given term t_i appearing in document $D_j \in D$, the *tf-idf* weighting is given as

$$tfidf(t_i, D_j) = count(t_i, D_j) \cdot log \frac{|D|}{|\{k : t_i \in D_k\}|} \tag{2}$$

This is the dot product of the term frequency (tf) and the inverse document frequency (idf), giving the name of the method. The term frequency, $count(t_i, D_j)$, is the number of times the term t_i appears in document D_j. The inverse document frequency, $log \frac{|D|}{|\{k:t_i \in D_k|\}}$, is the log of the inverse of the proportion of documents in the corpus containing the term. The resulting weights are used to update the matrix X given by the bag-of-words model.

The final step of the topic modelling process is to determine the distance between two vectors in X, giving the distance between the corresponding documents in terms of their topic. One such commonly used method is the cosine distance (Salton and McGill, 1986). While many other metrics exist, the cosine distance is often

[2] Topic modelling is also sometimes simply referred to as 'document classification' or 'document clustering'. We use the term 'topic modelling' here to differentiate between the categorisation of documents by topic from authorship.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

used for calculating the distance between documents in topic modelling applications (Aggarwal and Zhai, 2012).

3 Problem observation

In this section, we use a topic controlled authorship corpus to highlight the issues with alias matching and randomly separate aliases. We performed an experiment that compared the accuracy of randomly chosen sub-aliases compared to topic-separated sub-aliases using a topic controlled corpus. The corpus selected for this research was the AAAC 'A' corpus (Juola, 2004), in which thirteen students were asked to write essays on four topics. Each student has one essay on each topic, with one exception, giving fifty-one total documents with a mean document length of 4,541 characters and standard deviation of 1,070 characters. We chose two of these topics, and performed the experiment detailed below. We then repeated the experiment with combination of two topics, giving a total of six different subsets.

Each document was split into sub-documents of length 500 characters. This gave a larger corpus to draw sub-aliases from. Sub-aliases were then created either by sampling from the two topics at random ('random' sub-aliases) or splitting based on the topic of the documents ('topic' sub-aliases). We then used a large number of models to perform alias matching, using the procedure outlined in more detail below. For the random sub-aliases, we repeated the experiment for each model one hundred times, re-sampling to create new sub-aliases at each iteration. The expected result for this experiment was that the topic-based sub-aliases would produce significantly lower accuracies for alias matching than the random sub-aliases.

3.1 Results

We now present the results comparing random- and topic-based sub-aliases. As there was no variance in topic-based sub-aliases, only one iteration was performed. For random sub-aliases, one hundred iterations with re-sampled sub-aliases were chosen. However, for all tests, including the topic-based sub-aliases, ten-fold cross validation was performed and the mean results are taken as the quality of the algorithm.

The results from this experiment show a significant difference between the alias generation methods. Figure 2 highlights this result, with a drastic difference between the accuracy of random sub-aliases, which is very high, compared to topic-based sub-aliases, which is quite low. Models in this figure were ranked by their accuracy on random sub-aliases, highlighting the very large gap by the most accurate methods. Only in the worst performing 10% of methods did we find topic-based sub-alias accuracy was within the 95th percentile for the random sub-aliases. The topic-based sub-aliases results were similar to what many methods obtain on this dataset for classification accuracy (Layton *et al*., 2011c). We tested all topic combinations for this dataset and found that all repeated this combination. The mean accuracy values overall were over 0.75 for the random sub-aliases and under 0.14 for the topic sub-aliases. In all cases, the two tailed p-value for a t-test was less than e^{-170}, indicating

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

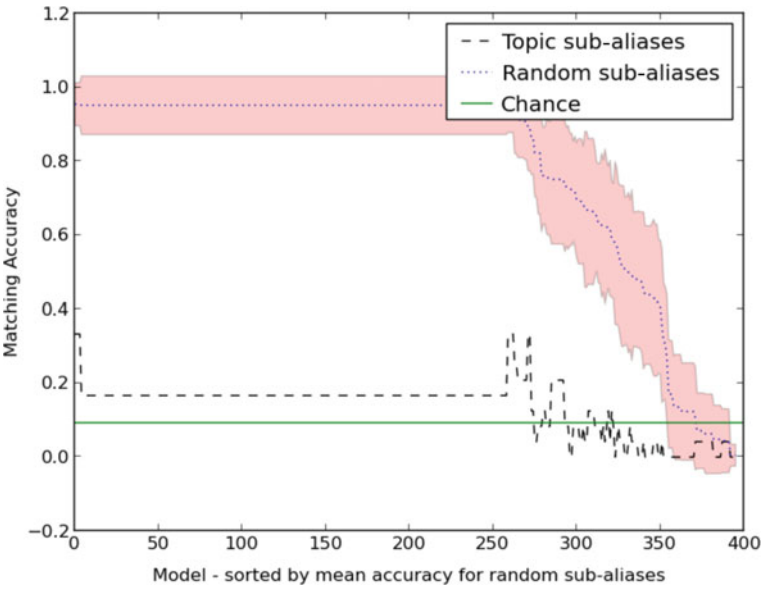


Fig. 2. (Colour online) Comparison between random sub-aliases and topic sub-aliases over all authorship models from this paper, for the AAAC A dataset. Shaded area is the 95th percentile for randomly chosen sub-aliases.

that the two distributions are highly unlikely to have been generated by the same distribution.

3.2 Effect of topic

To further show the effect of topic on sub-alias creation and subsequent alias matching, we repeated the experiment, but using sub-aliases with a mixture of sampling between topic and random. This mixture sub-alias was selected by choosing sub-documents from the above process with p% of sub-documents from the same topic and $(100 - p)$% from the opposite topic. When $p = 50$, we expected this procedure to have the same results as random sub-aliases. When $p = 100$ or when $p = 0$[3], we expected the same results as for the topic-based aliases above.

Figure 3 shows that this is indeed the case. When the probability is 50%, the prediction accuracy is at its highest and decreases when p both increases or decreases. As before, this same pattern was observed in all other topic combinations. A Pearson correlation of the relationship between the absolute distance to $p = 0.5$ and the accuracy was -0.956, $(p < 1e^{-27})$.

4 Methodology

First, LNG methods such as CNG are compared against the feature-based and bag-of-words approaches performed in previous research. In this experiment, a

[3] In this case, the sub-aliases are still topic-based, but simply reversed (topic 1 becomes sub-alias 2 and vice versa).

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

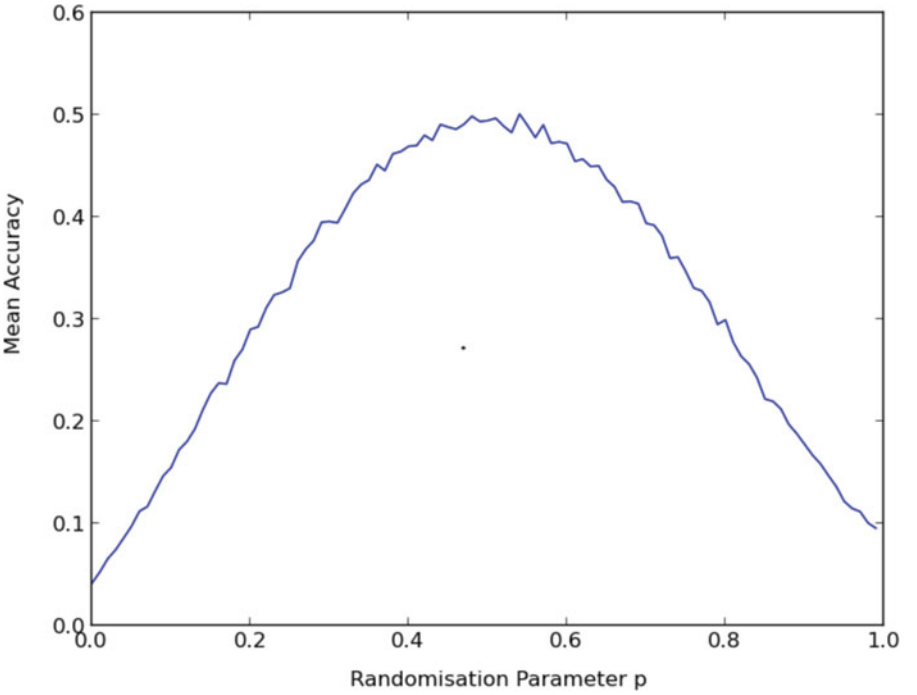


Fig. 3. (Colour online) Comparison between random sub-aliases and topic sub-aliases over many authorship models for AAAC A datasets. The randomisation parameter determines whether a document is selected from the same topic or the opposite topic, with 50% being completely randomly selected.

writer's documents are randomly split to form two sub-aliases. The algorithm then chooses the best matching sub-alias, aiming to discover the one originally belonging to the same author. We show that the LNG approaches generally outperform these previously tested methods. In the second experiment, each writer's documents are clustered into two sub-aliases using document topic clustering. The aim is to analyse the effect that topic has on the accuracy of the results. Previous research has suggested, without evidence, that character *n*-gram-based approaches are more susceptible to topic and context biases than other approaches to authorship analysis (Narayanan *et al*., 2012). We show that this is not true and that other methods are actually more susceptible than character-based approaches.

In this section, we first outline the datasets and algorithms used in both experiments. We then outline more fully how each of these experiments is performed and the method for analysing the results.

4.1 Datasets

Three datasets involving online communication were used in this research, since they are representative of real-world case studies for aliasing. The first two were taken from Twitter, an online micro-blogging website limiting posts to just 140 characters. The two datasets, Twitter Large and Twitter Small, were samples from the data

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

collected in Layton, Watters and Dazeley (2010). Twitter Large is the full set of posts from one hundred authors taken at random. Twitter Small is the first fifty postings by a separate set of one hundred randomly selected authors. In both cases, only aliases with at least ten documents were considered. The third dataset was taken from a dataset of blog postings from the online forum 'Chronicle of Higher Education' (Pillay and Solorio, 2010). While not limited in length, the blog postings are often small with around 400 to 600 characters per post. Only aliases with at least twenty postings were considered and of those aliases one hundred were chosen at random. A maximum of forty postings from each alias were taken.

Several preprocessing steps were undertaken for each dataset. For the Twitter-based datasets, we uniformly removed any tokens beginning with `@` and `#`, as these represent mentions of other users and topics, respectively. Both of these are out of the control of the author[4] and therefore removed. The forum data is considerably noisier than the Twitter data, including tags for formatting etc. In collecting the data, Pillay and Solorio removed such tags and mentions of the author/alias (Pillay and Solorio, 2010).

4.2 Tested authorship methods

There were two types of authorship analysis methods noted in Section 2: vector space and profile-based. Each was used in this experimentation with a number of parameters.

For vector space methods, we used each of the datasets identified by Narayanan *et al.* (2012): character, function words, syntactic and 'other'. While Narayanan *et al.* (2012) found that including all features provided the best accuracy in their tests, we considered all combinations of these datasets. We also tested the word features used by Novak *et al.* (2004), which were found to be most effective in their previous anti-aliasing experiments. For each sub-alias, the documents belonging to that sub-alias were represented in vector space form, then the vectors were averaged to form a centroid. The centroids were then compared using each of the Euclidean, cosine and correlation distance metrics. The sub-alias with the lowest distance to the candidate was chosen as the most likely match.

For profile-based methods, we considered all profile methods listed in Section 2: CNG, SCAP, RLP, d1, d2 and CNG-WPI. The parameters used for each of these methods were each combination of $n \in \{2, 3, 4, 5\}$ and $L \in \{100, 500, 1,000, 2,000, 5,000\}$. This range provided a representative coverage of the parameters used for these methods in previous research.

The KL divergence-based bag-of-words model employed by Novak *et al.* (2004) was also included in the tests. This model has two parameters; the number of words to include in the model (L) and the mixing parameter for including the corpus wide frequencies for smoothing α. We chose the same range for L as for the n-gram-based

[4] It could be argued that the topic hashtag is in the control of the user, as they select the topics they write about. However, it is very common for popular topics to dominate the use of this tag, which is why it was uniformly removed.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

models. Values for α range from 0 to 1 inclusive, in steps of 0.2. All combinations of L and α were tested.

4.3 Experiment 1: Randomly generated sub-aliases

In the first experiment, we evaluated different authorship analysis methods on their ability to correctly identify matching aliases for authors. This is performed for each of the datasets from Section 4.1 and the results analysed both on a overall-average scale and per-dataset.

For a given dataset, the documents from each of the authors were split into two sub-aliases. This was performed randomly, with the only constraint that each sub-alias had the same number of documents. Each model described in Section 4.2 was then evaluated using ten-fold cross validation (using the same sub-aliases for each fold). This procedure was run thirty times, each time generating new sub-aliases for all authors.

4.4 Experiment 2: Topic-based sub-aliases

The second experiment aimed to evaluate the effect of topic on the results. To do this, we repeated experiment 1 with sub-aliases chosen using topic clustering, rather than randomly. The documents were first stemmed, normalising word types. Any function words were discarded and then a bag-of-words model was applied to vectorise the documents. The documents for each author were clustered to create two evenly sized clusters with a high divergence by topic. In all other respects, this experiment was performed exactly as experiment 1. The expected result of this experiment was that the authorship methods would perform less effectively if they were susceptible to modelling topic and not authorship.

The first step of the process involved removing function words, which are often considered to include authorship information and are often discarded for the topic-based classification (Sebastiani, 2002). Secondly, the bag-of-words model was employed, collecting the frequency of all words appearing in at least 50% of all documents in each corpus. The most frequent 10,000 words fulfilling these criteria were kept for each corpus. As a third step, *tf-idf* weighting was applied. The above steps gave a matrix X such that $X_{i,j}$ was the *tf-idf* weighting of word i in document j. This methodology was performed using the topic modelling modules contained within scikits.learn version 0.11 (Pedregosa *et al*., 2011).

Finally, the documents from each author were clustered into two distinct clusters. A matrix D was created for each author, such that $D_{m,n}$ was the distance between documents m and n, with distance calculated using the cosine distance. The two documents with the highest distance were selected as exemplars. Each of the other documents was then ranked according to its minimum distance to one of the exemplars. Starting with the documents with the smallest distance, each was assigned to their nearest exemplar. This process continued until the number of documents assigned to either exemplar was more than 50% of the total number of documents. At this point, the rest of the documents were assigned to the *other*

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

exemplar, resulting in two distinct clusters that are approximately the same size. We used this method rather than an algorithm such as k-means used in other studies such as Layton *et al.* (2011a), to give evenly sized clusters. The k-means algorithm (and most other clustering algorithms) chooses clusters based on other criteria, often minimising some evaluation function that is not dependent on the size of the clusters.

As a further exercise, we replicated this second experiment leaving topic hashtags in the tweets for the topic clustering step and then removing them for the de-aliasing. We found only a small difference of generally less than 1% in accuracy for most methods and no discernible pattern to the results (i.e. one algorithm was not generally better performing than the other). For this reason, we conclude that the results were indicative of both including and excluding topic tags for the topic-based sub-alias creation.

As a final note, the documents were reverted to their original state for the authorship attribution. We did this to remove the effect of topic on the clusters, which may have led to overfitting. An example of this problem may be if an author pens a document in a topic not included in the training set, an overfitted model with topic features may incorrectly attribute this author due to the change in topic.

4.5 Experiment 3: Semantic topic-based sub-aliases

The bag-of-words model used in experiment 2 for sub-alias creation has the potential to positively affect character n-gram models. While word features are an obvious candidate for improvement based on the previous sub-alias creation, characters are closely linked to words and may also be affected positively. This may provide a positive benefit to those models that use character- or word-based features. To overcome this, we performed a final experiment that also incorporated semantic information. This experiment was performed exactly as experiment 2, with the exception that the semantic weighting described below was used, rather than the *tf-idf* weighting used in the previous experiments.

We employed the method used by Jing *et al.* (2006) that uses Wordnet to find semantically related terms. Wordnet has been used in a number of document (topic-based) clustering and classification (Watters and Patel, 1998; Hotho, Staab and Stumme, 2003; Sedding and Kazakov, 2004; Choudhury, Kimtani and Chakrabarty, 2012). The model used takes the input bag-of-words matrix, without the *tf-idf* weighting, and updates it based on the semantic relationship between words using the following equation:

$$\bar{x}_{j,i} = x_{j,i} + \sum_{k,i \neq k}^{m} \delta_{i,k} x_{j,k} \tag{3}$$

Where $x_{j,i}$ is the value of the ith term in the jth document and $\delta_{i,k}$ is the semantic distance between the ith and kth terms (a value between 0 and 1 with higher values indicating a higher semantic similarity). All rows in the matrix (i.e. the sum of weights for each document) were normalised to equal 1.0 for all documents.

The net effect of this update in weighting is that semantically related terms are increased in weight, even if they do not directly appear in the document. In an

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Table 1. *Abbreviations used in results tables*

Abbreviation	Meaning
CNG	Common *n*-grams method
SCAP	Source Code Author Profiles method
RLP	Recentred Local Profiles method
WPI	Weighted Profile Intersection method
d1	Stamatatos' d_1 method
d2	Stamatatos' d_2 method
C	Narayanan Character features
F	Narayanan Function word features
O	Narayanan's 'Other' features
KL-BOW	Kullback–Leibler with bag-of-words

example given in Jing *et al.* (2006), a document with the words 'ball' five times and 'basketball' three times results in a net value of 6.4 for the word 'football', even though it does not appear in the document at all. By using this semantic information, we drastically reduce the impact of characters on the bag-of-words representation, instead using word sense rather than word form.

5 Results and discussion

The results for the four experiments are given in this section. A large number of tests were performed over a large parameter range. In the interests of clarity and space considerations, some of the results have been summarised, but are available from the authors. The abbreviations used in the results table are shown in Table 1. For each experiment, we present two tables of results. The first table shows the top ten performing model/parameter combinations. As this list is usually dominated by the same form of model, we give a second table of each of the different types of model tested. In all cases, the best parameter combination is shown in this second table, for each of the model types. This gives a view of the comparative strength of each form of model, without listing each of the thousands of parameter combinations tested.

5.1 Experiment 1

The top results for experiment 1 were dominated by LNG methods, with CNG performing best overall and for each dataset. SCAP and RLP also performed favourably, which was expected since the methods are heavily related. Table 2 shows the top ten models (including parameters) and their performance on each corpus. Overall, there was no significant difference between relative corpora results, although it can be clearly seen that more data produces more accurate models. Including all tweets in the Large Twitter dataset performed better than the Small Twitter dataset, which in turn did better than the (generally) smaller Forum dataset, a pattern found in most authorship analysis studies.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Table 2. *Top ten models when sorted by mean accuracy for anti-aliasing for random sub-aliases. The maximum recorded score for each corpus is represented on this list. Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall mean
CNG $n = 5$ $L = 5{,}000$	0.9950	0.8600	0.3900	0.7483
SCAP $n = 5$ $L = 5{,}000$	0.9900	0.8400	0.3300	0.7200
CNG $n = 4$ $L = 5{,}000$	0.9950	0.8150	0.3450	0.7183
CNG $n = 3$ $L = 2{,}000$	0.9950	0.8000	0.3400	0.7117
CNG $n = 3$ $L = 5{,}000$	0.9950	0.8000	0.3400	0.7117
CNG $n = 5$ $L = 2{,}000$	0.9950	0.8600	0.2100	0.6883
RLP $n = 3$ $L = 5{,}000$	0.9750	0.6650	0.3900	0.6767
CNG $n = 5$ $L = 1{,}000$	0.9800	0.8600	0.1750	0.6717
SCAP n=5 L=2,000	0.9900	0.8400	0.1800	0.6700
RLP $n = 3$ $L = 2{,}000$	0.9700	0.6650	0.3750	0.6700

Table 3. *Top recorded accuracy for each model type (as well as variants for some models), for anti-aliasing for random sub-aliases (experiment 1). Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall mean
CNG $n = 5$ $L = 5{,}000$	0.9950	0.8600	0.3900	0.7483
SCAP $n = 5$ $L = 5{,}000$	0.9900	0.8400	0.3300	0.7200
RLP $n = 3$ $L = 5{,}000$	0.9750	0.6650	0.3900	0.6767
KL-BOW $L = 5{,}000$ $\alpha = 1.0$	0.8000	0.5050	0.1150	0.4733
Narayanan CF Euclidean	0.7900	0.3750	0.1200	0.4283
Narayanan C Euclidean	0.7550	0.3950	0.1300	0.4267
d1 $n = 2$ $L = 100$	0.7300	0.2750	0.1200	0.3750
WPI $n = 3$ $L = 1{,}000$	0.5950	0.4350	0.0750	0.3683
Narayanan F Euclidean	0.4900	0.1350	0.0450	0.2233
Narayanan CFO cosine	0.2800	0.1050	0.0250	0.1367
Narayanan CO correlation	0.2550	0.1000	0.0250	0.1267
Narayanan O correlation	0.2300	0.0950	0.0300	0.1183
Narayanan FO correlation	0.2250	0.0950	0.0250	0.1150
d2 $n = 2$ $L = 100$	0.1450	0.0250	0.0150	0.0617
Narayanan O Euclidean	0.0850	0.0450	0.0050	0.0450

In comparing the results of LNG-based models with other models, two aspects are clear. Firstly, there was a significant degradation in the results as less data appears, with accuracies for the large Twitter corpus being generally high, but the Forum corpus showing results barely above the chance rate of 0.05 for many model types. Secondly, KL-BOW performed comparably on the Twitter datasets to the better LNG methods, but scored quite poorly on the Forum dataset. One potential cause for this may be the less formal language in forums compared to tweets, resulting in fewer 'proper' words.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Table 4. *Top ten models when sorted by mean accuracy for anti-aliasing for the topic-separated sub-aliases (experiment 2). The maximum recorded score for each corpus is represented on this list. Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall mean
CNG $n = 5$ $L = 5{,}000$	0.8050	0.6100	0.2350	0.5500
CNG $n = 4$ $L = 5{,}000$	0.7650	0.5750	0.2150	0.5183
CNG $n = 5$ $L = 2{,}000$	0.8050	0.6100	0.1250	0.5133
CNG $n = 3$ $L = 5{,}000$	0.7750	0.5100	0.2100	0.4983
CNG $n = 5$ $L = 500$	0.6950	0.6100	0.1800	0.4950
CNG $n = 5$ $L = 1{,}000$	0.7000	0.6100	0.1500	0.4867
RLP $n = 3$ $L = 2{,}000$	0.6900	0.5050	0.2550	0.4833
CNG $n = 3$ $L = 2{,}000$	0.7750	0.5100	0.1450	0.4767
RLP $n = 3$ $L = 5{,}000$	0.7000	0.4650	0.2600	0.4750
CNG $n = 4$ $L = 2{,}000$	0.7300	0.5750	0.1100	0.4717

Table 5. *Top recorded accuracy for each model type (as well as variants for some models), for anti-aliasing for the topic-separated sub-aliases (experiment 2). Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall median
CNG $n = 5$ $L = 5{,}000$	0.8050	0.6100	0.2350	0.5500
RLP $n = 3$ $L = 2{,}000$	0.6900	0.5050	0.2550	0.4833
SCAP $n = 5$ $L = 5{,}000$	0.6700	0.5050	0.1250	0.4333
d1 $n = 2$ $L = 100$	0.5700	0.2600	0.0600	0.2967
WPI $n = 3$ $L = 1{,}000$	0.3850	0.3700	0.0750	0.2767
KL-BOW L = 5000 $\alpha = 1.0$	0.4400	0.2850	0.0600	0.2617
Narayanan C correlation	0.4300	0.1950	0.0650	0.2300
Narayanan CF Euclidean	0.3200	0.1600	0.0600	0.1800
Narayanan F Euclidean	0.2050	0.0800	0.0600	0.1150
Narayanan CO cosine	0.1000	0.0400	0.0050	0.0483
Narayanan CFO correlation	0.0950	0.0450	0.0050	0.0483
Narayanan FO correlation	0.0950	0.0350	0.0050	0.0450
Narayanan O correlation	0.1000	0.0300	0.0050	0.0450
d2 $n = 2$ $L = 100$	0.0350	0.0150	0.0600	0.0367
Narayanan O Euclidean	0.0650	0.0050	0.0000	0.0233

5.2 Experiment 2

Experiment 2, using topic-separated anti-aliasing experiment showed similar final results to experiment 1. LNG methods represented all of the top results, with CNG again performing best. The most significant result shown through this experiment was that all methods performed less for the topic-separated sub-aliases, compared to the randomly chosen sub-aliases. This difference was universal, but was more

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Table 6. *Percentage change between the results of experiment 1 and experiment 2. The best (least drop) of each model are listed and the median result is over all sub-results (not just summaries). Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall median
d2 $n = 4$ $L = 5{,}000$	0.0	−0.7500	0.0	−0.0
d1 $n = 2$ $L = 100$	−0.2192	−0.0545	−0.5000	−0.2089
RLP $n = 5$ $L = 100$	−0.2419	−0.2750	−0.1034	−0.2214
WPI $n = 3$ $L = 5{,}000$	−0.3646	−0.1494	0.1429	−0.2474
CNG $n = 5$ $L = 2{,}000$	−0.1910	−0.2907	−0.4048	−0.2542
KL-BOW $L = 100$ $\alpha = 0.32$	−0.0741	−0.4167	−0.8000	−0.3061
SCAP $n = 2$ $L = 100$	−0.3377	−0.3333	−0.6452	−0.3722
Narayanan C correlation	−0.4150	−0.4868	−0.4091	−0.4367

pronounced in some methods than others. Presented again are the top ten performing models, as well as the best performing of each model type.

5.3 Comparative results

Overall, the relative results between the two lists were similar. SCAP dropped significantly more than CNG or RLP, but still performed well compared to other methods. The feature-based approaches were again the worst performing methods, highlighting that these techniques are not suited to this application, at least for shorter documents tested here. The Spearman's rank coefficient between the two sets of results was 0.9883 with a given p-value of 0. The Pearson's correlation was similar at 0.9804 with a p-value of 0. The mean difference for all methods between the random and topic-based sub-aliases was 0.0868 and the mean difference as a percentage was 39.19%. For LNG-based methods, the difference was 35.15% (CNG = 30.16%, RLP = 32.15%, SCAP = 47.09%, d1 = 47.41%, d2 = 18.94%). For KL-BOW, the difference was 38.92%, while for the feature-based methods it was 55.31%.

Table 6 shows the percentage change between methods from experiment 1 to experiment 2. In all cases, the best (i.e. least drop in accuracy) was reported for each model type. While the vast majority of model/data pairs reported a drop in accuracy between experiment 1 and experiment 2, there were a few cases in which the accuracy improved. However, in all cases, the original models were poorly performing and this happened rarely enough to consider them to be chance events unlikely to consistently be repeated in real-world applications.

One significant result from the second experiment was that LNG methods tended to perform better comparatively *and relatively*, compared to non-LNG methods. Furthermore, the drop in accuracy was also less, as a proportion of accuracy in experiment 1. This is shown in Table 6, where LNG methods are near the top of the table for percentage of accuracy degradation. All changes in this table were statistically significant, as was the overall change in accuracy ($p < 1e^{20}$ in all cases).

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Table 7. *Top ten models when sorted by mean accuracy for anti-aliasing for semantically aware topic-separated sub-aliases (experiment 3). The maximum recorded score for each corpus is represented on this list. Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall mean
CNG $n = 5$ $L = 5{,}000$	0.7875	0.5612	0.1625	0.5037
CNG $n = 5$ $L = 2{,}000$	0.7177	0.5668	0.1077	0.4641
CNG $n = 4$ $L = 5{,}000$	0.7122	0.5185	0.1421	0.4576
CNG $n = 5$ $L = 1{,}000$	0.6508	0.5658	0.0909	0.4358
CNG $n = 4$ $L = 2{,}000$	0.6758	0.5227	0.0989	0.4325
SCAP $n = 5$ $L = 5{,}000$	0.7597	0.5017	0.0329	0.4314
CNG $n = 4$ $L = 1{,}000$	0.6418	0.5527	0.0822	0.4256
SCAP $n = 5$ $L = 2{,}000$	0.7135	0.5267	0.0304	0.4235
CNG $n = 3$ $L = 5{,}000$	0.6833	0.4440	0.1197	0.4157
CNG $n = 3$ $L = 2{,}000$	0.6782	0.4440	0.1102	0.4108

Table 8. *Top recorded accuracy for each model type (as well as variants for some models), for anti-aliasing for semantically aware topic-separated sub-aliases (experiment 3). Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall median
CNG $n = 5$ $L = 5{,}000$	0.7875	0.5612	0.1625	0.5037
SCAP $n = 5$ $L = 5{,}000$	0.7597	0.5017	0.0329	0.4314
RLP $n = 2$ $L = 1{,}000$	0.6742	0.3267	0.0803	0.3604
WPI $n = 2$ $L = 500$	0.6603	0.2655	0.0650	0.3303
d1 $n = 2$ $L = 100$	0.4843	0.2320	0.0299	0.2487
KLBOW $L = 500$ $\alpha = 0.98$	0.4082	0.2507	0.0494	0.2361
Narayanan C cosine	0.4297	0.1833	0.0342	0.2157
Narayanan CF cosine	0.3115	0.1500	0.0276	0.1630
Narayanan F Euclidean	0.1108	0.0580	0.0159	0.0616
Narayanan CFO correlation	0.1085	0.0590	0.0162	0.0612
Narayanan CO correlation	0.1060	0.0555	0.0162	0.0592
Narayanan FO correlation	0.0998	0.0540	0.0162	0.0567
Narayanan O correlation	0.0968	0.0517	0.0169	0.0551
d2 $n = 5$ $L = 100$	0.0282	0.0327	0.0033	0.0214

5.4 Experiment 3

The results from experiment 3, in which sub-aliases are created using a semantically aware topic modelling function, are given in Table 7 and show a similar pattern to experiment 2. The results showed a greater degradation, however, the overall rankings were relatively similar, especially to experiment 2. Furthermore, the per-model results shown in Table 8 show that there was little variation, despite the significant difference in the calculation of the profiles.

Finally, in Table 9, we compare the results of experiment 3 to experiment 1. Most notably, the word-based authorship did not reduce performance as before, since

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Table 9. *Percentage change between the results of experiment 1 and experiment 3. The best (least drop) of each model are listed and the median result is over all sub-results (not just summaries). Model abbreviations are listed in Table 1*

Model and parameters	Twitter Large	Twitter Small	Forum	Overall median
KL-BOW $L = 2{,}000$ $\alpha = 0.14$	−0.3190	0.1000	−0.5513	0.0000
WPI $n = 2$ $L = 500$	0.2006	−0.2414	−0.0005	−0.2500
d2 $n = 5$ $L = 100$	−0.6686	−0.1833	−0.6660	−0.2500
CNG $n = 4$ $L = 1{,}000$	−0.2869	−0.3219	−0.6087	−0.3221
d1 $n = 2$ $L = 100$	−0.3081	−0.3465	−0.6014	−0.3380
SCAP $n = 3$ $L = 100$	−0.3759	−0.3594	−0.7397	−0.3590
Narayanan CFO correlation	−0.6125	−0.4100	−0.3519	−0.4000
RLP $n = 5$ $L = 100$	−0.4909	−0.4642	−0.7351	−0.4250

words were not directly used for the sub-alias creation. A number of parameter sets had negligible results (fifteen from 256 combinations), with lower alpha values having less degradation in results. The LNG methods performed worse than previously, with the best LNG method still having a 25% degradation in results. The best Narayanan feature-based set had a loss of 40%. Surprisingly, the RLP method scored the 'worst best result' of 42.5% degradation, significantly worse than the previous comparison between experiment 1 and experiment 2. The overall degradation, excluding RLP, was comparable to the previous comparison of results.

6 Conclusions

In this paper, we highlighted that randomly chosen sub-aliases do not adequately represent the real-world problem of alias matching. In the theoretical sense, if an author is aiming to hide their identity by using multiple aliases, it is unlikely that the author will randomly choose an alias when posting a message. Instead, it is more likely that they would choose one alias for one purpose and another alias for a different one. In the practical sense, Section 3 showed that for a topic controlled dataset, using topic-based separation results in drastically reduced accuracy compared to random separation. Furthermore, we showed a correlation between the mixture of topic versus random and the final accuracy of the method. This correlation occurred across all methods we tested.

We then showed that using a topic-based clustering repeats this behaviour and can be used in datasets without topic identifiers. Thus, the methodology can be used in a wide variety of applications for alias matching. The use of topic-based sub-aliases allows for better evaluations of authorship methods for this task.

Furthermore, this method allowed us to test the efficacy of authorship methods, especially with respect to their overfitting of topic. If a model performed well with randomly chosen sub-aliases and poorly with topic-based sub-aliases, we suspect that the method is overfitting on topic more than author traits. In contrast, those that are high for both types of sub-aliases are better at modelling authorship. Our results for LNG methods are very high, with 99% accuracy achieved on the Large

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Twitter dataset for randomly selected sub-aliases. The accuracy does drop off with less data; the Small Twitter dataset recorded a high 84.5% and Forum just over 41%. Compared to the other results tested, LNG methods are highly effective both relatively and comparatively. This shows that the topic-based sub-aliases are more difficult to match than those randomly generated. Furthermore, it shows that LNG-based models are more robust to this degradation. An exception to this was RLP when using a semantically aware topic-based method.

This research also adds to the body of work that shows that LNG-based methods outperform feature-based approaches for authorship analysis applications. While it appears that current feature approaches are somewhat accurate in many settings, there is an ongoing trend that LNG-based methods perform better. With this research, the fears that character n-gram-based methods are topic-dependent appear to be unfounded. Further work on this result is needed to more accurately determine this conclusion across a larger range of authorship analysis applications.

Koppel, Schler and Argamon (2010) showed that by randomly selecting subsets of features for multiple classifiers and then ensembling the results, the accuracy of a character n-gram-based model (not an LNG method) can be drastically improved. As far as the authors are aware, this concept has not been tested in LNG models, although Layton *et al.* (2011a) did use an ensemble-based method with different n and L values. It would be worth investigating if this type of technique can be used to improve the performance of LNG methods for anti-aliasing, improving on results mentioned in this paper. Given that LNG methods appear to be more robust to topic-based sub-aliases, the ensemble decision of them may further improve robustness again. We also consider an application in non-language-based applications, such as the authorship analysis of compiled malware, using low-level features like API calls (Alazab *et al.*, 2010).

Acknowledgment

This research was conducted at the Internet Commerce Security Laboratory and was funded by the State Government of Victoria, IBM, Westpac, the Australian Federal Police and the University of Ballarat. More information can be found at *http://www.icsl.com.au*.

References

Aggarwal, C. C., and Zhai, C. X. (eds.) 2012. A survey of text classification algorithms. *Mining Text Data*, Springer, pp. 163–222. doi: 10.1007/978-1-4614-3223-4_6.

Alazab, M., Layton, R., Venkataraman, S., and Watters, P. 2010. Malware detection based on structural and behavioural features of API calls. In *Proceedings of the International Cyber Resilience Conference*, School of Computer and Information Science, Security Research Centre, Edith Cowan University, Perth, Western Australia.

Choudhury, J., Kimtani, D. K. and Chakrabarty, A. 2012. Text clustering using a WordNet-based knowledge-base and the Lesk Algorithm. *International Journal of Computer Applications* **48**(21): 20–4.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Clarke, R. V. G. 1997. *Situational Crime Prevention*. Guilderland, New York: Criminal Justice Press.

Escalante, H., Montes-y Gómez, M., and Solorio, T. 2011. A weighted profile intersection measure for profile-based authorship attribution. *Advances in Artificial Intelligence*, **7094**:232–43.

Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C. E., and Howald, B. S. 2007. Identifying authorship by byte-level n-grams: the source code author profile (SCAP) method. *International Journal of Digital Evidence* **6**(1): 1–18.

Holzer, R., Malin, B., and Sweeney, L. 2005. *Email Alias Detection Using Social Network Analysis*. PhD thesis. Information Networking Institute, Carnegie Mellon University.

Hotho, A., Staab, S., and Stumme, G. 2003. Ontologies improve text document clustering. In *Third IEEE International Conference on Data Mining, 2003. ICDM 2003*, Melbourne, Florida: IEEE, pp. 541–4.

Jing, L., Zhou, L., Ng, M. K., and Huang, J. Z. 2006. Ontology-based distance measure for text clustering. In *Proceedings of the Text Mining Workshop, SIAM International Conference on Data Mining*, Bethesda, Maryland.

Juola, P. 2004. Ad-hoc authorship attribution competition. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, Sweden, pp. 175–6.

Kešelj, V., Peng, F., Cercone, N., and Thomas, C. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*.

Koppel, M., Schler, J., and Argamon, S. 2010. Authorship attribution in the wild. *Language Resources and Evaluation* **45**(1): 83–94. ISSN 1574-020X. doi: 10.1007/s10579-009-9111-2.

Layton, R., McCombie, S., and Watters, P. A. 2012. Authorship attribution of IRC messages using inverse author frequency. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, Ballarat, Australia: IEEE, pp. 7–13.

Layton, R., and Watters, P. A. 2009. Determining provenance in phishing websites using automated conceptual analysis. In *eCrime Researchers Summit, 2009. eCRIME'09.*, Tacoma, WA, pp. 1–7.

Layton, R., Watters, P. A., and Dazeley, R. 2010. Authorship attribution for Twitter in 140 characters or less. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, Ballarat, Australia, pp. 1–8. ISBN 978-1-4244-8054-8. doi: 10.1109/CTC.2010.17.

Layton, R., Watters, P., and Dazeley, R. 2011a. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering* **1**(1): 1–26.

Layton, R., Watters, P., and Dazeley, R. 2011b. Automatically determining phishing campaigns using the USCAP methodology. In *eCrime Researchers Summit (eCrime), 2010*, Dallas, TX, pp. 1–8.

Layton, R., Watters, P. A., and Dazeley, R. 2011c. Recentred local profiles for authorship attribution. *Journal of Natural Language Engineering* **18**(3): 293–312. doi: 10.1017/S1351324911000180. Available on CJO 2011.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Luyckx, K., and Daelemans, W. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing* **26**(1): 35.

Narayanan, A., Paskov, H., Gong, N. Z., and Bethencourt, J. 2012. On the feasibility of internet-scale author identification. In *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*, San Francisco, CA,.

Novak, J., Raghavan, P., and Tomkins, A. 2004. Anti-aliasing on the web. In *Proceedings of the 13th conference on World Wide Web - WWW '04*, New York: ACM, pp. 30–9. doi: 10.1145/988672.988678.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: machine learning in Python . *Journal of Machine Learning Research* **12**: 2825–30.

Pillay, S. R., and Solorio, T. 2010. Authorship attribution of web forum posts. In *eCrime Researchers Summit (eCrime), 2010*, Dallas, TX, pp. 1–7.

Rudman, J. 1998. The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities* **31**: 351–65.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* **24**(5): 513–23.

Salton, G., and McGill, M. J. 1986. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Schein, A. I., Caver, J. F., Honaker, R. J., and Martell, C. H. 2010. Author attribution evaluation with novel topic cross-validation. In *KDIR*, Valencia, Spain, pp. 206–15.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* **34**(1): 1–47.

Sedding, J., and Kazakov, D. 2004. Wordnet-based text document clustering. In *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*, Geneva: Association for Computational Linguistics, pp. 104–13.

Solorio, T., Pillay, S., Raghavan, S., and Montes-y Gómez, M. 2011. Modality specific meta features for authorship attribution in web forum posts. In *IJCNLP*, Chiang Mai, Thailand, pp. 156–64.

Stabek, A., Watters, P. A., and Layton, R. 2010. The seven scam types: mapping the terrain of cybercrime. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*, Ballarat, Australia, pp. 41–51.

Stamatatos, E. 2007. Author identification using imbalanced and limited training texts. In *18th International Workshop on Database and Expert Systems Applications, 2007. DEXA'07.* , Regensburg, pp. 237–41.

Ureche, O., Layton, R., and Watters, P. A. 2012. Towards an implementation of information flow security using semantic web technologies. In *2012 Third Cybercrime and Trustworthy Computing Workshop*, Ballarat, Australia, pp. 1–8.

Watters, P. A., McCombie, S., Layton, R., and Pieprzyk, J. 2012. Characterising and predicting cyber attacks using the Cyber Attacker Model Profile (CAMP). *Journal of Money Laundering Control* **15**(4): 430–41.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press

Watters, P. A., and Patel, M. 1998. Modeling lexical-semantic processes using wordnet. *Glot International* **3**(9–10): 23–4.

Zheng, R., Li, J., Chen, H., and Huang, Z. 2005. A framework for authorship identification of online messages: writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* **57**: 378–93.

https://doi.org/10.1017/S1351324913000272 Published online by Cambridge University Press