Authorship Attribution with Latent Dirichlet Allocation

Yanir Seroussi **Ingrid Zukerman** **Fabian Bohnert**
Faculty of Information Technology, Monash University
Clayton, Victoria 3800, Australia
firstname.lastname@monash.edu

Abstract

The problem of authorship attribution – attributing texts to their original authors – has been an active research area since the end of the 19th century, attracting increased interest in the last decade. Most of the work on authorship attribution focuses on scenarios with only a few candidate authors, but recently considered cases with tens to thousands of candidate authors were found to be much more challenging. In this paper, we propose ways of employing Latent Dirichlet Allocation in authorship attribution. We show that our approach yields state-of-the-art performance for both a few and many candidate authors, in cases where these authors wrote enough texts to be modelled effectively.

1 Introduction

The problem of authorship attribution – attributing texts to their original authors – has received considerable attention in the last decade (Juola, 2006; Stamatatos, 2009). Most of the work in this field focuses on cases where texts must be attributed to one of a few candidate authors, e.g., (Mosteller and Wallace, 1964; Gamon, 2004). Recently, researchers have turned their attention to scenarios with tens to thousands of candidate authors (Koppel et al., 2011). In this paper, we study authorship attribution with few to many candidate authors, and introduce a new method that achieves state-of-the-art performance in the latter case.

Our approach to authorship attribution consists of building models of authors and their texts using *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003). We compare these models to models built from texts with unknown authors to find the most likely authors of these texts (Section 3.2). Our evaluation shows that our approach yields a higher accuracy than the method recently introduced by Koppel et al. (2011) in several cases where prolific authors are considered, while requiring less runtime (Section 4).

This paper is structured as follows. Related work is surveyed in Section 2. Our LDA-based approach to authorship attribution is described in Section 3, together with the baselines we considered in our evaluation. Section 4 presents and discusses the results of our evaluation, and Section 5 discusses our conclusions and plans for future work.

2 Related Work

The field of authorship attribution predates modern computing. For example, in the late 19th century, Mendenhall (1887) suggested that word length can be used to distinguish works by different authors. In recent years, increased interest in authorship attribution was fuelled by advances in machine learning, information retrieval, and natural language processing (Juola, 2006; Stamatatos, 2009).

Commonly used features in authorship attribution range from "shallow" features, such as token and character n-gram frequencies, to features that require deeper analysis, such as part-of-speech and rewrite rule frequencies (Stamatatos, 2009). As in other text classification tasks, *Support Vector Machines* (SVMs) have delivered high accuracy, as they are designed to handle feature vectors of high dimensionality (Juola, 2006). For example, *one-vs.-all* (OVA) is an effective approach to using binary SVMs for multi-class (i.e., multi-author) problems (Rifkin and Klautau, 2004). Given A authors,

Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 181–189,
Portland, Oregon, USA, 23–24 June 2011.

OVA trains A binary classifiers, where each classifier is trained on texts by one author as positive examples and all the other texts as negative examples. However, if A is large, each classifier has many more negative than positive examples, often yielding poor results due to class imbalance (Raskutti and Kowalczyk, 2004). Other setups, such as *one-vs.-one* or *directed acyclic graph*, require training $O(A^2)$ classifiers, making them impractical where thousands of authors exist. *Multi-class SVMs* have also been suggested, but they generally perform comparably to OVA while taking longer to train (Rifkin and Klautau, 2004). Hence, using SVMs for scenarios with many candidate authors is problematic (Koppel et al., 2011). Recent approaches to employing binary SVMs consider class similarity to improve performance (Bickerstaffe and Zukerman, 2010; Cheng et al., 2007). We leave experiments with such approaches for future work (Section 5).

In this paper, we focus on authorship attribution with many candidate authors. This problem was previously addressed by Madigan et al. (2005) and Luyckx and Daelemans (2008), who worked on datasets with texts by 114 and 145 authors respectively. In both cases, the reported results were much poorer than those reported in the binary case. More recently, Koppel et al. (2011) considered author similarity to handle cases with thousands of candidate authors. Their method, which we use as our baseline, is described in Section 3.1.

Our approach to authorship attribution utilises *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) to build models of authors from their texts. LDA is a generative probabilistic model that is traditionally used to find topics in textual data. The main idea behind LDA is that each document in a corpus is generated from a distribution of topics, and each word in the document is generated according to the per-topic word distribution. Blei et al. (2003) showed that using LDA for dimensionality reduction can improve performance for supervised text classification. We know of only one case where LDA was used in authorship attribution: Rajkumar et al. (2009) reported preliminary results on using LDA topic distributions as feature vectors for SVMs, but they did not compare the results obtained with LDA-based SVMs to those obtained with SVMs trained on tokens directly. Our comparison shows that both methods perform comparably (Section 4.3).

Nonetheless, the main focus of our work is on authorship attribution with *many* candidate authors, where it is problematic to use SVMs. Our *LDA+Hellinger* approach employs LDA *without* SVM training (Section 3.2), yielding state-of-the-art performance in several scenarios (Section 4).

3 Authorship Attribution Methods

This section describes the authorship attribution methods considered in this paper. While all these methods can employ various representations of documents, e.g., token frequencies or part-of-speech n-gram frequencies, we only experimented with token frequencies.[1] This is because they are simple to extract, and can achieve good performance (Section 4). Further, the focus of this paper is on comparing the performance of our methods to that of the baseline methods. Thus, we leave experiments on other feature types for future work (Section 5).

3.1 Baselines

We consider two baseline methods, depending on whether there are two or many candidate authors. If there are only two, we use *Support Vector Machines* (SVMs), which have been shown to deliver state-of-the-art performance on this task (Juola, 2006). If there are many, we follow Koppel et al.'s (2011) approach, which we denote *KOP*.

The main idea behind KOP is that different pairs of authors may be distinguished by different subsets of the feature space. Hence, KOP randomly chooses k_1 subsets of size $k_2 F$ ($k_2 < 1$) from a set of F features; for each of the k_1 subsets, it calculates the cosine similarity between a test document and all the documents by one author (each author is represented by one feature vector); it then outputs the author who had most of the top matches. KOP also includes a threshold σ^* to handle cases where a higher level of precision is required, at the cost of lower recall. If the top-matching author was the top match less than σ^* times, then KOP outputs "unknown author". In our experiments we set $\sigma^* = 0$ to obtain full coverage, as this makes it easier to interpret the results using a single measure of accuracy.

[1] Token frequency is the token count divided by the total number of tokens.

3.2 Authorship Attribution with LDA

In this work, we follow the extended LDA model defined by Griffiths and Steyvers (2004). Under the assumptions of the extended model, given a corpus of M documents, a document i with N tokens is generated by choosing a document topic distribution $\theta_i \sim \text{Dir}(\alpha)$, where $\text{Dir}(\alpha)$ is a T-dimensional symmetric Dirichlet distribution, and α and T are parameters of the model. Then, each token in the document w_{ij} is generated by choosing a topic from the document topic distribution $z_{ij} \sim \text{Multinomial}(\theta_i)$, and choosing a token from the token topic distribution $w_{ij} \sim \text{Multinomial}(\phi_{z_{ij}})$, where $\phi_{z_{ij}} \sim \text{Dir}(\beta)$, and β is a parameter of the model. The model can be inferred from the data using Gibbs sampling, as outlined in (Griffiths and Steyvers, 2004) – an approach we follow in our experiments.

Note that the topics obtained by LDA do not have to correspond to actual, human-interpretable topics. A more appropriate name may be "latent factors", but we adopt the convention of calling these factors "topics" throughout this paper. The meaning of the factors depends on the type of tokens that are used as input to the LDA inference process. For example, if stopwords are removed from the corpus, the resulting factors often, but not necessarily, correspond to topics. However, if only stopwords are retained, as is commonly done in authorship attribution studies, the resulting factors lose their interpretability as topics; rather, they can be seen as stylistic markers. Note that even if stopwords are discarded, nothing forces the factors to stand for actual topics. Indeed, in a preliminary experiment on a corpus of movie reviews and message board posts, we found that some factors correspond to topics, with words such as "noir" and "detective" considered to be highly probable for one topic. However, other factors seemed to correspond to authorship style as reflected by authors' vocabulary, with netspeak words such as "wanna", "alot" and "haha" assigned to one topic, and words such as "compelling" and "beautifully" assigned to a different topic.

We consider two ways of using LDA in authorship attribution: (1) *Topic SVM*, and (2) *LDA+Hellinger*. The LDA part of both approaches consists of applying a frequency filter to the features in the training documents,[2] and then using LDA to reduce the dimensionality of each document to a topic distribution of dimensionality T.

Topic SVM. The topic distributions are used as features for a binary SVM classifier that discriminates between authors. This approach has been employed in the past for document classification, e.g., in (Blei et al., 2003), but it has been applied to authorship attribution only in a limited study that considered just stopwords (Rajkumar et al., 2009). In Section 4.3, we present the results of more thorough experiments in applying this approach to binary authorship attribution. Our results show that the performance of this approach is comparable to that obtained without using LDA. This indicates that we do not lose authorship-related information when employing LDA, even though the dimensionality of the document representations is greatly reduced.

LDA+Hellinger. This method is our main contribution, as it achieves state-of-the-art performance in authorship attribution with many candidate authors, where it is problematic to use SVMs (Section 2).

The main idea of our approach is to use the *Hellinger distance* between document topic distributions to find the most likely author of a document:[3] $D(\theta_1, \theta_2) = \sqrt{\frac{1}{2}\sum_{t=1}^{T}\left(\sqrt{\theta_{1,t}} - \sqrt{\theta_{2,t}}\right)^2}$ where θ_i is a T-dimensional multinomial topic distribution, and $\theta_{i,t}$ is the probability of the t-th topic.

We propose two representations of an author's documents: *multi-document* and *single-document*.

- *Multi-document (LDAH-M).* The LDA model is built based on all the training documents. Given a test document, we measure the Hellinger distance between its topic distribution and the topic distributions of the training documents. The author with the lowest mean distance for all of his/her documents is returned as the most likely author of the test document.

[2] We employed frequency filtering because it has been shown to be a scalable and effective feature selection method for authorship attribution tasks (Stamatatos, 2009). We leave experiments with other feature selection methods for future work.

[3] We considered other measures for comparing topic distributions, including Kullback-Leibler divergence and Bhattacharyya distance. From these measures, only Hellinger distance satisfies all required properties of a distance metric. Hence, we used Hellinger distance.

- *Single-document (LDAH-S).* Each author's documents are concatenated into a single document (the *profile document*), and the LDA model is learned from the profile documents.[4] Given a test document, the Hellinger distance between the topic distributions of the test document and all the profile documents is measured, and the author of the profile document with the shortest distance is returned.

The time it takes to learn the LDA model depends on the number of Gibbs samples S, the number of tokens in the training corpus W, and the number of topics T. For each Gibbs sample, the algorithm iterates through all the tokens in the corpus, and for each token it iterates through all the topics. Thus, the time complexity of learning the model is $O(SWT)$. Once the model is learned, inferring the topic distribution of a test document of length N takes $O(SNT)$. Therefore, the time it takes to classify a document when using LDAH-S is $O(SNT+AT)$, where A is the number of authors, and $O(T)$ is the time complexity of calculating the Hellinger distance between two T-dimensional distributions. The time it takes to classify a document when using LDAH-M is $O(SNT + MT)$, where M is the total number of training documents, and $M \geq A$, because every candidate author has written at least one document.

An advantage of LDAH-S over LDAH-M is that LDAH-S requires much less time to classify a test document when many documents per author are available. However, this improvement in runtime may come at the price of accuracy, as authorship markers that are present only in a few short documents by one author may lose their prominence if these documents are concatenated to longer documents. In our evaluation we found that LDAH-M outperforms LDAH-S when applied to one of the datasets (Section 4.3), while LDAH-S yields a higher accuracy when applied to the other two datasets (Sections 4.4 and 4.5). Hence, we present the results obtained with both variants.

[4]Concatenating all the author documents into one document has been named the *profile-based* approach in previous studies, in contrast to the *instance-based* approach, where each document is considered separately (Stamatatos, 2009).

4 Evaluation

In this section, we describe the experimental setup and datasets used in our experiments, followed by the evaluation of our methods. We evaluate Topic SVM for binary authorship attribution, and LDA+Hellinger on a binary dataset, a dataset with tens of authors, and a dataset with thousands of authors. Our results show that LDA+Hellinger yields a higher accuracy than Koppel et al.'s (2011) baseline method in several cases where prolific authors are considered, while requiring less runtime.

4.1 Experimental Setup

In all the experiments, we perform ten-fold cross validation, employing stratified sampling where possible. The results are evaluated using classification accuracy, i.e., the percentage of test documents that were correctly assigned to their author. Note that we use different accuracy ranges in the figures that present our results for clarity of presentation. Statistically significant differences are reported when $p < 0.05$ according to a paired two-tailed t-test.

We used the LDA implementation from LingPipe (`alias-i.com/lingpipe`) and the SVM implementation from Weka (`www.cs.waikato.ac.nz/ml/weka`). Since our focus is on testing the impact of LDA, we used a linear SVM kernel and the default SVM settings. For the LDA parameters, we followed Griffiths and Steyvers (2004) and the recommendations in LingPipe's documentation, and set the Dirichlet hyperparameters to $\alpha = \min(0.1, 50/T)$ and $\beta = 0.01$, varying only the number of topics T. We ran the Gibbs sampling process for $S = 1000$ iterations, and based the document representations on the last sample. While taking more than one sample is generally considered good practice (Steyvers and Griffiths, 2007), we found that the impact of taking several samples on accuracy is minimal, but it substantially increases the runtime. Hence, we decided to use only one sample in our experiments.

4.2 Datasets

We considered three datasets that cover different writing styles and settings: *Judgement*, *IMDb62* and *Blog*. Table 1 shows a summary of these datasets.

The **Judgement dataset** contains judgements by three judges who served on the Australian High

	Judgement	IMDb62	Blog
Authors	3	62	19,320
Texts	1,342	62,000	678,161
Texts per Author	Dixon: 902 McTiernan: 253 Rich: 187	1,000	Mean: 35.10 Stddev.: 104.99

Table 1: Dataset Statistics

Court from 1913 to 1975: Dixon, McTiernan and Rich (available for download from `www.csse.monash.edu.au/research/umnl/data`). In this paper, we considered the Dixon/McTiernan and the Dixon/Rich binary classification cases, using judgements from non-overlapping periods (Dixon's 1929–1964 judgements, McTiernan's 1965–1975, and Rich's 1913–1928). We removed numbers from the texts to ensure that dates could not be used to discriminate between judges. We also removed quotes to ensure that the classifiers take into account only the actual author's language use.[5] Employing this dataset in our experiments allows us to test our methods on formal texts with a minimal amount of noise.

The **IMDb62 dataset** contains 62,000 movie reviews by 62 prolific users of the Internet Movie database (IMDb, `www.imdb.com`, available upon request from the authors of (Seroussi et al., 2010)). Each user wrote 1,000 reviews. This dataset is noisier than the Judgement dataset, since it may contain spelling and grammatical errors, and the reviews are not as professionally edited as judgements. This dataset allows us to test our approach in a setting where all the texts have similar themes, and the number of authors is relatively small, but is already much larger than the number of authors considered in traditional authorship attribution settings.

The **Blog dataset** is the largest dataset we considered, containing 678,161 blog posts by 19,320 authors (Schler et al., 2006) (available for download from `u.cs.biu.ac.il/~koppel`). In contrast to IMDb reviews, blog posts can be about any topic, but the large number of authors ensures that every topic is likely to interest at least some authors. Koppel et al. (2011) used a different blog dataset consisting of 10,240 authors in their work on authorship attribution with many candidate authors. Unfortunately, their dataset is not publicly available. However, authorship attribution is more challenging on the dataset we used, because they imposed some restrictions on their dataset, such as setting a minimal number of words per author, and truncating the training and testing texts so that they all have the same length. The dataset we use has no such restrictions.

[5]We removed numbers and quotes by matching regular expressions for numbers and text in quotation marks, respectively.

4.3 LDA in Binary Authorship Attribution

In this section, we present the results of our experiments with the Judgement dataset (Section 4.2), testing the use of LDA in producing feature vectors for SVMs and the performance of our LDA+Hellinger methods (Section 3.2).

In all the experiments, we employed a classifier ensemble to address the class imbalance problem present in the Judgement dataset, which contains 5 times more texts by Dixon than by Rich, and over 3 times more texts by Dixon than by McTiernan (Table 1). Dixon's texts are randomly split into 5 or 3 subsets, depending on the other author (Rich or McTiernan respectively), and the base classifiers are trained on each subset of Dixon's texts together with all the texts by the other judge. Given a text by an unknown author, the classifier outputs are combined using majority voting. We found that the accuracies obtained with an ensemble are higher than those obtained with a single classifier. We did not require the vote to be unanimous, even though this increases precision, because we wanted to ensure full coverage of the test dataset. This enables us to compare different methods using only an accuracy measure.[6]

Experiment 1. Figure 1 shows the results of an experiment that compares the accuracy obtained using SVMs with token frequencies as features (*Token SVMs*) with that obtained using LDA topic distributions as features (*Topic SVMs*). We experimented with several filters on token frequency, and different numbers of LDA topics (5, 10, 25, 50, . . ., 250). The x-axis labels describe the frequency filters: the minimum and maximum token frequencies, and the approximate number of unique tokens left after filtering (in thousands). We present only the results obtained with 10, 25, 100 and 200 topics, as the re-

[6]For all our experiments, the results for the Dixon/McTiernan case are comparable to those for Dixon/Rich. Therefore, we omit the Dixon/McTiernan results to conserve space.

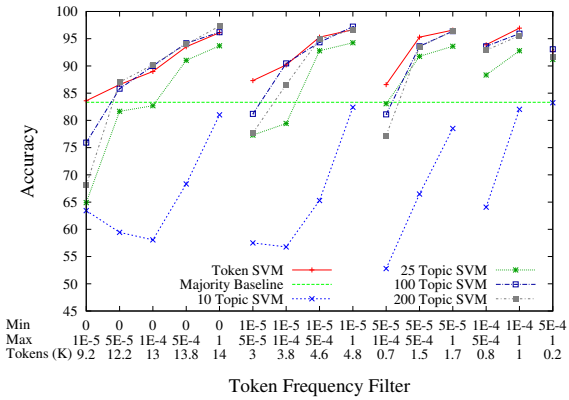


Figure 1: LDA Features for SVMs in Binary Authorship Attribution (Judgement dataset, Dixon/Rich)

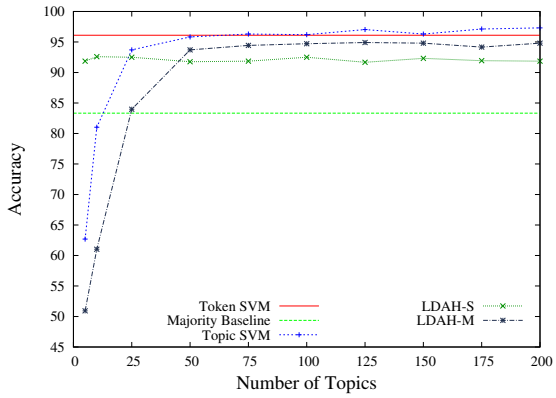


Figure 2: LDA+Hellinger in Binary Authorship Attribution (Judgement dataset, Dixon/Rich)

sults obtained with other topic numbers are consistent with the presented results, and the results obtained with 225 and 250 topics are comparable to the results obtained with 200 topics.

Our results show that setting a maximum bound on token frequency filters out important authorship markers, regardless of whether LDA is used or not (performance drops). This shows that it is unlikely that discriminative LDA topics correspond to actual topics, as the most frequent tokens are mostly non-topical (e.g., punctuation and function words).

An additional conclusion is that using LDA for feature reduction yields results that are comparable to those obtained using tokens directly. While Topic SVMs seem to perform slightly better than Token SVMs, the differences between the best results obtained with the two approaches are not statistically significant. However, the number of features that the SVMs consider when topics are used is usually much smaller than when tokens are used directly, especially when no token filters are used (i.e., when the minimum frequency is 0 and the maximum frequency is 1). This makes it easy to apply LDA to different datasets, since the token filtering parameters may be domain-dependent, and LDA yields good results without filtering tokens.

Experiment 2. Figure 2 shows the results of an experiment that compares the performance of the single profile document (LDAH-S) and multiple author documents (LDAH-M) variants of our LDA+Hellinger approach to the results obtained with Token SVMs and Topic SVMs. As in Experiment 1, we employ classifier ensembles, where the base classifiers are either SVMs or LDA+Hellinger classifiers. We did not filter tokens, since Experiment 1 indicates that filtering has no advantage over not filtering tokens. Instead, Figure 2 presents the accuracy as a function of the number of topics.

Note that we did not expect LDA+Hellinger to outperform SVMs, since LDA+Hellinger does not consider inter-class relationships. Indeed, Figure 2 shows that this is the case (the differences between the best Topic SVM results and the best LDAH-M results are statistically significant). However, LDA+Hellinger still delivers results that are much better than the majority baseline (the differences between LDA+Hellinger and the majority baseline are statistically significant). This leads us to hypothesise that LDA+Hellinger will perform well in cases where it is problematic to use SVMs due to the large number of candidate authors. We verify this hypothesis in the following sections.

One notable result is that LDAH-S delivers high accuracy even when only a few topics are used, while LDAH-M requires about 50 topics to outperform LDAH-S (all the differences between LDAH-S and LDAH-M are statistically significant). This may be because there are only two authors, so LDAH-S builds the LDA model based only on two profile documents. Hence, even 5 topics are enough to obtain two topic distributions that are sufficiently different to discriminate the authors' test documents. The reason LDAH-M outperforms LDAH-S when more topics are considered may be that some important authorship markers lose their prominence in the profile documents created by LDAH-S.

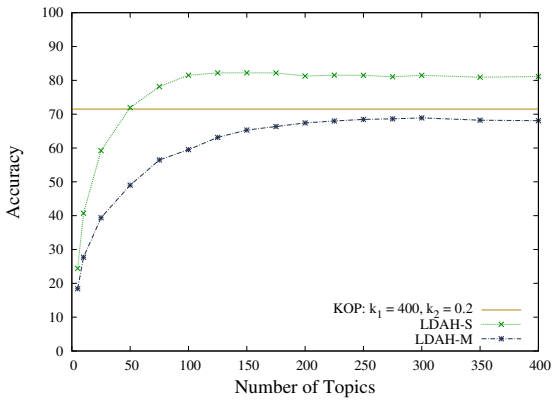


Figure 3: LDA+Hellinger with Tens of Authors (IMDb62 dataset)

4.4 LDA+Hellinger with Tens of Authors

In this section, we apply our LDA+Hellinger approaches to the IMDb62 dataset (Section 4.2), and compare the obtained results to those obtained with Koppel et al.'s (2011) method (KOP). To this effect, we first established a KOP best-performance baseline by performing parameter tuning experiments for KOP. Figure 3 shows the results of the comparison of the accuracies obtained with our LDA+Hellinger methods to the best accuracy yielded by KOP (obtained in the parameter tuning experiment).

For this experiment, we ran our LDA+Hellinger variants with 5, 10, 25, 50, ..., 300, 350 and 400 topics. The highest LDAH-M accuracy was obtained with 300 topics (Figure 3). However, LDAH-S yielded a much higher accuracy than LDAH-M. This may be because the large number of training texts per author (900) may be too noisy for LDAH-M. That is, the differences between individual texts by each author may be too large to yield a meaningful representation of the author if they are considered separately. Finally, LDAH-S requires only 50 topics to outperform KOP, and outperforms KOP by about 15% for 150 topics. All the differences between the methods are statistically significant.

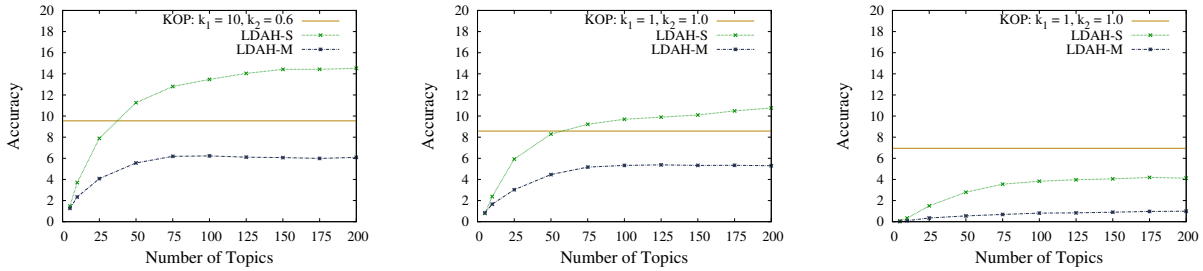
This experiment shows that LDAH-S models the authors in IMDb62 more accurately than KOP. The large improvement in accuracy shows that the compact author representation employed by LDAH-S, which requires only 150 topics to obtain the highest accuracy, has more power to discriminate between authors than KOP's much heavier representation, of 400 subsets with more than 30,000 features each. In addition, the *per-fold* runtime of the KOP baseline was 93 hours, while LDAH-S required only 15 hours per fold to obtain the highest accuracy.

4.5 LDA+Hellinger with Thousands of Authors

In this section, we compare the performance of our LDA+Hellinger variants to the performance of KOP on several subsets of the Blog dataset (Section 4.2). For this purpose, we split the dataset according to the prolificness of the authors, i.e., we ordered the authors by the number of blog posts, and considered subsets that contain all the posts by the 1000, 2000, 5000 and 19320 most prolific authors.[7] Due to the large number of posts, we could not run KOP for more than $k_1 = 10$ iterations on the smallest subset of the dataset and 5 iterations on the other subsets, as the runtime was prohibitive for more iterations. For example, 10 iterations on the smallest subset required about 90 hours per fold (the LDA+Hellinger runtimes were substantially shorter, with maximum runtimes of 56 hours for LDAH-S and 77 hours for LDAH-M, when 200 topics were considered). Interestingly, running KOP for 5 iterations on the larger subsets decreased performance compared to running it for 1 iteration. Thus, on the larger subsets, the most accurate KOP results took less time to obtain than those of our LDA+Hellinger variants.

Figure 4 shows the results of this experiment. For each author subset, it compares the results obtained by LDAH-S and LDAH-M to the best result obtained by KOP. All the differences between the methods are statistically significant. For up to 2000 prolific authors (Figures 4(a), 4(b)), LDAH-S outperforms KOP by up to 50%. For 5000 prolific users (figure omitted due to space limitations), the methods perform comparably, and KOP outperforms LDAH-S by a small margin. However, with all the authors (Figure 4(c)), KOP yields a higher accuracy than both LDA+Hellinger variants. This may be because considering non-prolific authors introduces noise that results in an LDA model that does not capture the differences between authors. However, it is encouraging that LDAH-S outperforms KOP when less than 5000 prolific authors are considered.

[7] These authors make up about 5%, 10%, 25% and exactly 100% of the authors, but they wrote about 50%, 65%, 80% and exactly 100% of the texts, respectively.



(a) 1,000 Prolific Authors (b) 2,000 Prolific Authors (c) 19,320 (all) Authors

Figure 4: LDA+Hellinger with Thousands of Authors (Blog dataset)

The accuracies obtained in this section are rather low compared to those obtained in the previous sections. This is not surprising, since the authorship attribution problem is much more challenging with thousands of candidate authors. This challenge motivated the introduction of the σ^* threshold in KOP (Section 3.1). Our LDA+Hellinger variants can also be extended to include a threshold: if the Hellinger distance of the best-matching author is greater than the threshold, the LDA+Hellinger algorithm would return "unknown author". We leave experiments with this extension to future work, as our focus in this paper is on comparing LDA+Hellinger to KOP, and we believe that this comparison is clearer when no thresholds are used.

5 Conclusions and Future Work

In this paper, we introduced an approach to authorship attribution that models texts and authors using Latent Dirichlet Allocation (LDA), and considers the distance between the LDA-based representations of the training and test texts when classifying test texts. We showed that our approach yields state-of-the-art performance in terms of classification accuracy when tens or a few thousand authors are considered, and prolific authors exist in the training data. This accuracy improvement was achieved together with a substantial reduction in runtime compared to Koppel et al.'s (2011) baseline method.

While we found that our approach performs well on texts by prolific authors, there is still room for improvement on authors who have not written many texts – an issue that we will address in the future. One approach that may improve performance on such authors involves considering other types of features than tokens, such as parts of speech and character n-grams. Since our approach is based on LDA, it can easily employ different feature types, which makes this a straightforward extension to the work presented in this paper.

In the future, we also plan to explore ways of extending LDA to model authors directly, rather than using it as a black box. Authors were considered by Rosen-Zvi et al. (2004; 2010), who extended LDA to form an author-topic model. However, this model was not used for authorship attribution, and was mostly aimed at topic modelling of multi-authored texts, such as research papers.

Another possible research direction is to improve the scalability of our methods. Our approach, like Koppel et al.'s (2011) baseline, requires linear time in the number of possible authors to classify a single document. One possible way of reducing the time needed for prediction is by employing a hierarchical approach that builds a tree of classifiers based on class similarity, as done by Bickerstaffe and Zukerman (2010) for the sentiment analysis task. Under this framework, class similarity (in our case, author similarity) can be measured using LDA, while small groups of classes can be discriminated using SVMs.

In addition to authorship attribution, we plan to employ text-based author models in user modelling tasks such as rating prediction – a direction that we already started working on by successfully applying our LDA-based approach to model users for the rating prediction task (Seroussi et al., 2011).

Acknowledgements

This research was supported in part by grant LP0883416 from the Australian Research Council. The authors thank Russell Smyth for the collaboration on initial results on the judgement dataset.

References

Adrian Bickerstaffe and Ingrid Zukerman. 2010. A hierarchical classifier applied to multi-way sentiment detection. In *COLING 2010: Proceedings of the 23rd International Conference on Computational Linguistics*, pages 62–70, Beijing, China.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Haibin Cheng, Pang-Ning Tan, and Rong Jin. 2007. Localized support vector machine and its efficient algorithm. In *SDM 2007: Proceedings of the 7th SIAM International Conference on Data Mining*, pages 461–466, Minneapolis, MN, USA.

Michael Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 611–617, Geneva, Switzerland.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235.

Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *COLING 2008: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 513–520, Manchester, UK.

David Madigan, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. In *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, St. Louis, MO, USA.

Thomas C. Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214S):237–246.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Arun Rajkumar, Saradha Ravi, Venkatasubramanian Suresh, M. Narasimha Murthy, and C. E. Veni Madhavan. 2009. Stopwords and stylometry: A latent Dirichlet allocation approach. In *Proceedings of the NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond (Poster Session)*, Whistler, BC, Canada.

Bhavani Raskutti and Adam Kowalczyk. 2004. Extreme re-balancing for SVMs: A case study. *ACM SIGKDD Explorations Newsletter*, 6(1):60–69.

Ryan Rifkin and Aldebaro Klautau. 2004. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5(Jan):101–141.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI 2004: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, Banff, AB, Canada.

Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. 2010. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, 28(1):1–38.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–205, Stanford, CA, USA.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2010. Collaborative inference of sentiments from texts. In *UMAP 2010: Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization*, pages 195–206, Waikoloa, HI, USA.

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2011. Personalised rating prediction for new users using latent factor models. In *Hypertext 2011: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, Eindhoven, The Netherlands.

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

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors, *Handbook of Latent Semantic Analysis*, pages 427–448. Lawrence Erlbaum Associates.