Authorship Attribution with Author-aware Topic Models

Yanir Seroussi **Fabian Bohnert** **Ingrid Zukerman**

Faculty of Information Technology, Monash University
Clayton, Victoria 3800, Australia
firstname.lastname@monash.edu

Abstract

Authorship attribution deals with identifying the authors of anonymous texts. Building on our earlier finding that the Latent Dirichlet Allocation (LDA) topic model can be used to improve authorship attribution accuracy, we show that employing a previously-suggested Author-Topic (AT) model outperforms LDA when applied to scenarios with many authors. In addition, we define a model that combines LDA and AT by representing authors and documents over two disjoint topic sets, and show that our model outperforms LDA, AT and support vector machines on datasets with many authors.

1 Introduction

Authorship attribution (AA) has attracted much attention due to its many applications in, e.g., computer forensics, criminal law, military intelligence, and humanities research (Stamatatos, 2009). The traditional problem, which is the focus of our work, is to attribute *test texts* of unknown authorship to one of a set of known authors, whose *training texts* are supplied in advance (i.e., a supervised classification problem). While most of the early work on AA focused on formal texts with only a few possible authors, researchers have recently turned their attention to informal texts and tens to thousands of authors (Koppel et al., 2011). In parallel, topic models have gained popularity as a means of analysing such large text corpora (Blei, 2012). In (Seroussi et al., 2011), we showed that methods based on *Latent Dirichlet Allocation* (LDA) – a popular topic model by Blei et al. (2003) – yield good AA performance. However, LDA does not model authors explicitly, and we are not aware of any previous studies that apply *author-aware* topic models to traditional AA. This paper aims to address this gap.

In addition to being the first (to the best of our knowledge) to apply Rosen-Zvi et al.'s (2004) *Author-Topic Model* (AT) to traditional AA, the main contribution of this paper is our *Disjoint Author-Document Topic Model* (DADT), which addresses AT's limitations in the context of AA. We show that DADT outperforms AT, LDA, and linear support vector machines on AA with many authors.

2 Disjoint Author-Document Topic Model

Background. Our definition of DADT is motivated by the observation that when authors write texts on the same issue, specific words must be used (e.g., texts about LDA are likely to contain the words "topic" and "prior"), while other words vary in frequency according to author style. Also, texts by the same author share similar style markers, independently of content (Koppel et al., 2009). DADT aims to separate *document words* from *author words* by generating them from two disjoint topic sets of $T^{(D)}$ *document topics* and $T^{(A)}$ *author topics*.

Lacoste-Julien et al. (2008) and Ramage et al. (2009) (among others) also used disjoint topic sets to represent document labels, and Chemudugunta et al. (2006) separated corpus-level topics from document-specific words. However, we are unaware of any applications of these ideas to AA. The closest work we know of is by Mimno and McCallum (2008), whose DMR model outperformed AT in AA

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 264–269,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

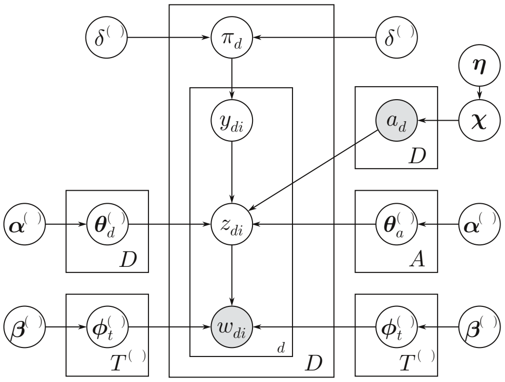
$\delta^{( )}$
π_d
$\delta^{( )}$
$\boldsymbol{\eta}$
y_{di}
a_d
$\boldsymbol{\chi}$
D
$\boldsymbol{\alpha}^{( )}$
$\boldsymbol{\theta}_d^{( )}$
z_{di}
$\boldsymbol{\theta}_a^{( )}$
$\boldsymbol{\alpha}^{( )}$
D
A
$\boldsymbol{\beta}^{( )}$
$\boldsymbol{\phi}_t^{( )}$
w_{di}
$\boldsymbol{\phi}_t^{( )}$
$\boldsymbol{\beta}^{( )}$
d
$T^{( )}$
D
$T^{( )}$

Figure 1: The Disjoint Author-Document Topic Model

of *multi-authored* texts (DMR does not use disjoint topic sets). We use AT rather than DMR, since we found that AT outperforms DMR in AA of *single-authored* texts, which are the focus of this paper.

The Model. Figure 1 shows DADT's graphical representation, with document-related parameters on the left (the LDA component), and author-related parameters on the right (the AT component). We define the model for single-authored texts, but it can be easily extended to multi-authored texts.

The generative process for DADT is described below. We use $\mathcal{D}$ and $\mathcal{C}$ to denote the Dirichlet and categorical distributions respectively, and A, D and V to denote the number of authors, documents, and unique vocabulary words respectively. In addition, we mark each step as coming from either **L**DA or **A**T, or as new in **D**ADT.

Global level:

- **L.** For each document topic t, draw a word distribution $\boldsymbol{\phi}_t^{(D)} \sim \mathcal{D}\left(\boldsymbol{\beta}^{(D)}\right)$, where $\boldsymbol{\beta}^{(D)}$ is a length-V vector.
- **A.** For each author topic t, draw a word distribution $\boldsymbol{\phi}_t^{(A)} \sim \mathcal{D}\left(\boldsymbol{\beta}^{(A)}\right)$, where $\boldsymbol{\beta}^{(A)}$ is a length-V vector.
- **A.** For each author a, draw the author topic distribution $\boldsymbol{\theta}_a^{(A)} \sim \mathcal{D}\left(\boldsymbol{\alpha}^{(A)}\right)$, where $\boldsymbol{\alpha}^{(A)}$ is a length-$T^{(A)}$ vector.
- **D.** Draw a distribution over authors $\boldsymbol{\chi} \sim \mathcal{D}(\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a length-A vector.

Document level: For each document d:

- **L.** Draw d's topic distribution $\boldsymbol{\theta}_d^{(D)} \sim \mathcal{D}\left(\boldsymbol{\alpha}^{(D)}\right)$, where $\boldsymbol{\alpha}^{(D)}$ is a length-$T^{(D)}$ vector.
- **D.** Draw d's author $a_d \sim \mathcal{C}(\boldsymbol{\chi})$.
- **D.** Draw d's topic ratio $\pi_d \sim \mathrm{Beta}\left(\delta^{(A)}, \delta^{(D)}\right)$, where $\delta^{(A)}$ and $\delta^{(D)}$ are scalars.

Word level: For each word index i in document d:

- **D.** Draw di's topic indicator $y_{di} \sim \mathrm{Bernoulli}(\pi_d)$.
- **L.** If $y_{di} = 0$, draw a *document* topic $z_{di} \sim \mathcal{C}\left(\boldsymbol{\theta}_d^{(D)}\right)$ and word $w_{di} \sim \mathcal{C}\left(\boldsymbol{\phi}_{z_{di}}^{(D)}\right)$.
- **A.** If $y_{di} = 1$, draw an *author* topic $z_{di} \sim \mathcal{C}\left(\boldsymbol{\theta}_{a_d}^{(A)}\right)$ and word $w_{di} \sim \mathcal{C}\left(\boldsymbol{\phi}_{z_{di}}^{(A)}\right)$.

DADT versus AT. DADT might seem similar to AT with "fictitious" authors, as described by Rosen-Zvi et al. (2010) (i.e., AT trained with an additional unique "fictitious" author for each document, allowing it to adapt to individual documents and not only to authors). However, there are several key differences between DADT and AT.

First, in DADT *author topics are disjoint from document topics*, with different priors for each topic set. Thus, the number of author topics can be different from the number of document topics, enabling us to vary the number of author topics according to the number of authors in the corpus.

Second, DADT *places different priors on the word distributions* for author topics and document topics ($\boldsymbol{\beta}^{(A)}$ and $\boldsymbol{\beta}^{(D)}$ respectively). Stopwords are known to be strong indicators of authorship (Koppel et al., 2009), and DADT allows us to use this knowledge by assigning higher weights to the elements of $\boldsymbol{\beta}^{(A)}$ that correspond to stopwords than to such elements in $\boldsymbol{\beta}^{(D)}$.

Third, DADT *learns the ratio between document words and author words* on a per-document basis, and makes it possible to specify a prior belief of what this ratio should be. We found that specifying a prior belief that about 80% of each document is composed of author words yielded better results than using AT's approach, which evenly splits each document into author and document words.

Fourth, DADT *defines the process that generates authors*. This allows us to consider the number of texts by each author when performing AA. This also enables the potential use of DADT in a semi-supervised setting by training on unlabelled texts, which we plan to explore in the future.

3 Authorship Attribution Methods

We experimented with the following AA methods, using token frequency features, which are good predictors of authorship (Koppel et al., 2009).

Baseline: Support Vector Machines (SVMs). Koppel et al. (2009) showed that SVMs yield good AA performance. We use linear SVMs in a one-versus-all setup, as implemented in LIBLINEAR (Fan et al., 2008), reporting results obtained with the best cost parameter values.

Baseline: LDA + Hellinger (LDA-H). This approach uses the Hellinger distances of topic distributions to assign test texts to the closest author. In (Seroussi et al., 2011), we experimented with two variants: (1) each author's texts are concatenated before building the LDA model; and (2) no concatenation is performed. We found that the latter approach performs poorly in cases with many candidate authors. Hence, we use only the former approach in this paper. Note that when dealing with single-authored texts, concatenating each author's texts yields an LDA model that is equivalent to AT.

AT. Given an inferred AT model (Rosen-Zvi et al., 2004), we calculate the probability of the test text words for each author a, assuming it was written by a, and return the most probable author. We do not know of any other studies that used AT in this manner for single-authored AA. We expect this method to outperform LDA-H as it employs AT directly, rather than relying on an external distance measure.

AT-FA. Same as AT, but built with an additional unique "fictitious" author for each document.

DADT. Given our DADT model, we assume that the test text was written by a "new" author, and infer this author's topic distribution, the author/document topic ratio, and the document topic distribution. We then calculate the probability of each author given the model's parameters, the test text words, and the inferred author/document topic ratio and document topic distribution. The most probable author is returned. We use this method to avoid inferring the document-dependent parameters separately for each author, which is infeasible when many authors exist. A version that marginalises over these parameters will be explored in future work.

4 Evaluation

We compare the performance of the methods on two publicly-available datasets: (1) *PAN'11*: emails with 72 authors (Argamon and Juola, 2011); and (2) *Blog:* blogs with 19,320 authors (Schler et al., 2006). These datasets represent realistic scenarios of AA of user-generated texts with many candidate authors. For example, Chaski (2005) notes a case where an employee who was terminated for sending a racist email claimed that any person with access to his computer could have sent the email.

Experimental Setup. Experiments on the PAN'11 dataset followed the setup of the PAN'11 competition (Argamon and Juola, 2011): We trained all the methods on the given training subset, tuned the parameters according to the results on the given validation subset, and ran the tuned methods on the given testing subset. In the Blog experiments, we used tenfold cross validation as in (Seroussi et al., 2011).

We used collapsed Gibbs sampling to train all the topic models (Griffiths and Steyvers, 2004), running 4 chains with a burn-in of 1,000 iterations. In the PAN'11 experiments, we retained 8 samples per chain with spacing of 100 iterations. In the Blog experiments, we retained 1 sample per chain due to runtime constraints. Since we cannot average topic distribution estimates obtained from training samples due to topic exchangeability (Steyvers and Griffiths, 2007), we averaged the distances and probabilities calculated from the retained samples. For test text sampling, we used a burn-in of 100 iterations and averaged the parameter estimates over the next 100 iterations in a similar manner to Rosen-Zvi et al. (2010). We found that these settings yield stable results across different random seed values.

We found that the number of topics has a larger impact on accuracy than other configurable parameters. Hence, we used symmetric topic priors, setting all the elements of $\boldsymbol{\alpha}^{(D)}$ and $\boldsymbol{\alpha}^{(A)}$ to $\min\{0.1, 5/T^{(D)}\}$ and $\min\{0.1, 5/T^{(A)}\}$ respectively.[1] For all models, we set $\beta_w = 0.01$ for each word w as the base measure for the prior of words in topics. Since DADT allows us to encode our prior knowledge that stopword use is indicative of authorship, we set $\beta_w^{(D)} = 0.01 - \epsilon$ and $\beta_w^{(A)} = 0.01 + \epsilon$ for all w, where w is a stopword.[2] We set $\epsilon = 0.009$, which improved accuracy by up to one percentage point over using $\epsilon = 0$. Finally, we set $\delta^{(A)} = 4.889$ and $\delta^{(D)} = 1.222$ for DADT. This encodes our prior

[1] We tested Wallach et al.'s (2009) method of obtaining asymmetric priors, but found that it did not improve accuracy.

[2] We used the stopword list from www.lextek.com/manuals/onix/stopwords2.html.

Method	PAN'11 Validation	PAN'11 Testing	Blog Prolific	Blog Full
SVM	48.61%	53.31%	33.31%	24.13%
LDA-H	34.95%	42.62%	21.61%	7.94%
AT	46.68%	53.08%	37.56%	23.03%
AT-FA	20.68%	24.23%	—	—
DADT	**54.24%**	**59.08%**	**42.51%**	**27.63%**

Table 1: Experiment results

belief that 0.8 ± 0.15 of each document is composed of author words. We found that this yields better results than an uninformed uniform prior of $\delta^{(A)} = \delta^{(D)} = 1$ (Seroussi et al., 2012). In addition, we set $\eta_a = 1$ for each author a, yielding smoothed estimates for the corpus distribution of authors χ.

To fairly compare the topic-based methods, we used the same overall number of topics for all the topic models. We present only the results obtained with the best topic settings: 100 for PAN'11 and 400 for Blog, with DADT's author/document topic splits being 90/10 for PAN'11, and 390/10 for Blog. These splits allow DADT to de-noise the author representations by allocating document words to a relatively small number of document topics. It is worth noting that AT can be seen as an extreme version of DADT, where all the topics are author topics. A future extension is to learn the topic balance automatically, e.g., in a similar manner to Teh et al.'s (2006) method of inferring the number of topics in LDA.

Results. Table 1 shows the results of our experiments in terms of classification accuracy (i.e., the percentage of test texts correctly attributed to their author). The PAN'11 results are shown for the validation and testing subsets, and the Blog results are shown for a subset containing the 1,000 most prolific authors and for the full dataset of 19,320 authors.

Our DADT model yielded the best results in all cases (the differences between DADT and the other methods are statistically significant according to a paired two-tailed t-test with $p < 0.05$). We attribute DADT's superior performance to the de-noising effect of the disjoint topic sets, which appear to yield author representations of higher predictive quality than those of the other models.

As expected, AT significantly outperformed LDA-H. On the other hand, AT-FA performed much worse than all the other methods on PAN'11, probably because of the inherent noisiness in using the same topics to model both authors and documents. Hence, we did not run AT-FA on the Blog dataset.

DADT's PAN'11 testing result is close to the third-best accuracy from the PAN'11 competition (Argamon and Juola, 2011). However, to the best of our knowledge, DADT obtained the best accuracy for a fully-supervised method that uses only unigram features. Specifically, Kourtis and Stamatatos (2011), who obtained the highest accuracy (65.8%), assumed that all the test texts are given to the classifier at the same time, and used this additional information with a semi-supervised method; while Kern et al. (2011) and Tanguy et al. (2011), who obtained the second-best (64.2%) and third-best (59.4%) accuracies respectively, used various feature types (e.g., features obtained from parse trees). Further, preprocessing differences make it hard to compare the methods on a level playing field. Nonetheless, we note that extending DADT to enable semi-supervised classification and additional feature types are promising future work directions.

While all the methods yielded relatively low accuracies on Blog due to its size, topic-based methods were more strongly affected than SVM by the transition from the 1,000 author subset to the full dataset. This is probably because topic-based methods use a single model, making them more sensitive to corpus size than SVM's one-versus-all setup that uses one model per author. Notably, an oracle that chooses the correct answer between SVM and DADT when they disagree yields an accuracy of 37.15% on the full dataset, suggesting it is worthwhile to explore ensembles that combine the outputs of SVM and DADT (we tried using DADT topics as additional SVM features, but this did not outperform DADT).

5 Conclusion

This paper demonstrated the utility of using author-aware topic models for AA: AT outperformed LDA, and our DADT model outperformed LDA, AT and SVMs in cases with noisy texts and many authors. We hope that these results will inspire further research into the application of topic models to AA.

Acknowledgements

This research was supported in part by Australian Research Council grant LP0883416. We thank Mark Carman for fruitful discussions on topic modelling.

References

Shlomo Argamon and Patrick Juola. 2011. Overview of the international authorship identification competition at PAN-2011. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers)*, Amsterdam, The Netherlands.

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

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Carole E. Chaski. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1).

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS 2006: Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pages 241–248, Vancouver, BC, Canada.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874.

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

Roman Kern, Christin Seifert, Mario Zechner, and Michael Granitzer. 2011. Vote/veto meta-classifier for authorship identification. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers)*, Amsterdam, The Netherlands.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

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

Ioannis Kourtis and Efstathios Stamatatos. 2011. Author identification using semi-supervised learning. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers)*, Amsterdam, The Netherlands.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS 2008: Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 897–904, Vancouver, BC, Canada.

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI 2008: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 411–418, Helsinki, Finland.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP 2009: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore.

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. 2011. Authorship attribution with latent Dirichlet allocation. In *CoNLL 2011: Proceedings of the 15th International Conference on Computational Natural Language Learning*, pages 181–189, Portland, OR, USA.

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2012. Authorship attribution with author-aware topic models. Technical Report 2012/268, Faculty of Information Technology, Monash University, Clayton, VIC, Australia.

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.

Ludovic Tanguy, Assaf Urieli, Basilio Calderone, Nabil Hathout, and Franck Sajous. 2011. A multitude of linguistically-rich features for authorship attribution. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers)*, Amsterdam, The Netherlands.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *NIPS 2009: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, pages 1973–1981, Vancouver, BC, Canada.