Authorship Attribution with Topic Models

Yanir Seroussi*
Monash University

Ingrid Zukerman**
Monash University

Fabian Bohnert†
Monash University

Authorship attribution deals with identifying the authors of anonymous texts. Traditionally, research in this field has focused on formal texts, such as essays and novels, but recently more attention has been given to texts generated by on-line users, such as e-mails and blogs. Authorship attribution of such on-line texts is a more challenging task than traditional authorship attribution, because such texts tend to be short, and the number of candidate authors is often larger than in traditional settings. We address this challenge by using topic models to obtain author representations. In addition to exploring novel ways of applying two popular topic models to this task, we test our new model that projects authors and documents to two disjoint topic spaces. Utilizing our model in authorship attribution yields state-of-the-art performance on several data sets, containing either formal texts written by a few authors or informal texts generated by tens to thousands of on-line users. We also present experimental results that demonstrate the applicability of topical author representations to two other problems: inferring the sentiment polarity of texts, and predicting the ratings that users would give to items such as movies.

1. Introduction

Authorship attribution has attracted much attention due to its many applications in, for example, computer forensics, criminal law, military intelligence, and humanities research (Juola 2006; Stamatatos 2009; Argamon and Juola 2011). The traditional problem, which is the focus of this article, is to attribute anonymous **test texts** to one of a set of known candidate authors, whose **training texts** are supplied in advance (i.e., supervised classification). Whereas most of the early work on authorship attribution focused on formal texts with only a few candidate authors, researchers have recently

* Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia. E-mail: yanir.seroussi@monash.edu.
** Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia. E-mail: ingrid.zukerman@monash.edu.
† Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia. E-mail: fabian.bohnert@monash.edu.

Submission received: 30 December 2012; revised submission received: 9 May 2013; accepted for publication: 23 June 2013.

doi:10.1162/COLI_a_00173

turned their attention to scenarios involving informal texts and tens to thousands of authors (Koppel, Schler, and Argamon 2011; Luyckx and Daelemans 2011). In parallel, topic models have gained popularity as a means of discovering themes in such large text corpora (Blei 2012). This article explores authorship attribution with topic models, extending the work presented by Seroussi and colleagues (Seroussi, Zukerman, and Bohnert 2011; Seroussi, Bohnert, and Zukerman 2012) by reporting additional experimental results and applications of topic-based author representations that go beyond traditional authorship attribution.

Topic models work by defining a probabilistic representation of the latent structure of corpora through latent factors called **topics**, which are commonly associated with distributions over words (Blei 2012). For example, in the popular Latent Dirichlet Allocation (LDA) topic model, each document is associated with a distribution over topics, and each word in the document is generated according to its topic's distribution over words (Blei, Ng, and Jordan 2003). The word distributions often correspond to a human-interpretable notion of topics, but this is not guaranteed, as interpretability depends on the corpus used for training the model. Indeed, when we ran LDA on a data set of movie reviews and message board posts, we found that some word distributions 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. This finding motivated our use of LDA for authorship attribution (Seroussi, Zukerman, and Bohnert 2011).

One limitation of LDA is that it does not model authors explicitly. This led us to use Rosen-Zvi et al.'s (2004) Author-Topic (AT) model to obtain improved authorship attribution results (Seroussi, Bohnert, and Zukerman 2012). However, AT is also limited in that it does not model documents. We addressed this limitation through the Disjoint Author-Document Topic (DADT) model—a topic model that draws on the strengths of LDA and AT, while addressing their limitations by integrating them into a single model. Our DADT model extends the model introduced by Seroussi, Bohnert, and Zukerman (2012), which could only be trained on single-authored texts. In this article, we provide a detailed account of the extended model. In addition, we offer experimental results for five data sets, extending the results by Seroussi, Bohnert, and Zukerman (2012), which were restricted to two data sets of informal texts with many authors. Our experiments show that DADT yields state-of-the-art performance on these data sets, which contain either formal texts written by a few authors or informal texts where the number of candidate authors ranges from 62 to about 20,000.

Although our evaluation is focused on single-authored texts, AT and DADT can also be used to model authors based on multi-authored texts, such as research papers. To demonstrate the potential utility of this capability of the models, we present the results of a preliminary study, where we use AT and DADT to identify anonymous reviewers based on publicly available information (reviewer lists and the reviewers' publications, which are often multi-authored). Our results indicate that reviewers may be identified with moderate accuracy, at least in small conference tracks and workshops. We hope that these results will help fuel discussions on the issue of reviewer anonymity.

Our finding that topic models yield good authorship attribution performance indicates that they capture aspects of authorship style, which is known to be indicative of author characteristics such as demographic information and personality traits (Argamon et al. 2009). This is in addition to the well-established result that topic models can be used to represent authors' interests (Rosen-Zvi et al. 2004). An implication of these results is that topic models may be used to obtain text-based representations of users in scenarios where user-generated texts are available. We demonstrate this by

showing how topic models can be utilized to improve the performance of methods we developed to address the popular tasks of polarity inference and rating prediction.

This article is structured as follows. Section 2 surveys related work. Section 3 discusses LDA, AT, and DADT and the author representations they yield. Section 4 introduces authorship attribution methods, which are evaluated in Section 5. Section 6 presents applications of our topic-based approach, and Section 7 concludes the article.

2. Related Work

Authorship attribution has a long history that predates modern computing. For example, Mendenhall (1887) suggested that word length can be used to distinguish works by different authors. Modern interest in authorship attribution is commonly traced back to Mosteller and Wallace's (1964) study on applying Bayesian statistical analysis of function word frequencies to uncover the authors of the *Federalist Papers* (Juola 2006; Koppel, Schler, and Argamon 2009; Stamatatos 2009). The interest in authorship attribution is due to its many applications in areas such as computer forensics, criminal law, military intelligence, and humanities research. In recent years, authorship attribution research has been fuelled by advances in natural language processing, text mining, machine learning, information retrieval, and statistical analysis. This has motivated the organization of workshops and competitions to facilitate the development and comparison of authorship attribution methods (Juola 2004; Argamon and Juola 2011).

In this article, we focus on the **closed-set** attribution task, where training texts by the candidate authors are supplied in advance, and for each test text, the goal is to attribute the text to the correct author out of the candidate authors (Argamon and Juola 2011). Related tasks include **open-set** attribution, where some test texts may not have been written by any of the candidate authors, and **verification**, where texts by only one candidate author are supplied in advance, and the task is to verify whether test texts were written by the candidate author (Koppel and Schler 2004; Sanderson and Guenter 2006; Koppel, Schler, and Argamon 2011).

Regardless of the task, a challenge currently faced by researchers in the field is addressing scenarios with many candidate authors and varying amounts of data per author (Argamon and Juola 2011; Koppel, Schler, and Argamon 2011; Luyckx and Daelemans 2011). This challenge is illustrated by the corpus chosen for the PAN'11 competition (Argamon and Juola 2011), which contains short e-mails by tens of authors. Other examples are Koppel, Schler, and Argamon's (2011) work on a corpus of blog posts by thousands of authors, and Luyckx and Daelemans's (2011) study of the effect of the number of authors and training set size on authorship attribution performance on data sets of student essays. Our approach to authorship attribution addresses this challenge by using topic models, which are known to successfully deal with varying amounts of text (Blei 2012).

We know of only one previous case where topic models were used for authorship attribution of single-authored texts: Rajkumar et al. (2009) reported preliminary results on using LDA topic distributions as feature vectors for support vector machines (SVM), but they did not compare the results obtained with LDA-based SVM to those obtained with SVM trained on tokens only (we present the results of such a comparison in Section 5). We know of two related studies that followed the publication of our initial LDA-based results (Seroussi, Zukerman, and Bohnert 2011): Wong, Dras and Johnson's (2011) work on native language identification with LDA, and Pearl and Steyvers's (2012) study of authorship verification where some of the features are topic distributions. Although Wong, Dras and Johnson reported only limited success (perhaps

because an author's native language may manifest itself in only a few words, or maybe due to data-set-specific issues), Pearl and Steyvers found that topical representations helped them achieve state-of-the-art verification accuracy. Pearl and Steyvers's findings further strengthen our hypothesis that topic models yield meaningful author representations. We take this observation one step further by defining our DADT model, and applying it to several authorship attribution scenarios, where it yields better performance than LDA-based approaches and methods based on the AT model (Section 5).

A line of research that has garnered much interest in recent years is the definition of **generic** topic models that incorporate metadata labels (Blei 2012). These models can be divided into two types: **upstream** models, which use the labels to constrain the topics, and **downstream** models, which generate the labels from the topics (Mimno and McCallum 2008). Generic models have the appealing advantage of obviating the need to define a new model for each new task (e.g., they may be used to obtain author representations by defining a metadata label for each author). However, this advantage may come at the price of increased computational complexity or poorer performance than that of task-specific models (Mimno and McCallum 2008). As the focus of our work is on modeling authors, we experimented only with LDA and with the task-specific topic models discussed in Section 3 (AT and DADT, which model authors explicitly). The applicability of generic models to authorship attribution is an open question that would be interesting to investigate in the future. Nonetheless, most of the generic models surveyed here have properties that make them unsuitable for our purposes.

Examples of generic upstream models include DiscLDA (Lacoste-Julien, Sha, and Jordan 2008), Labeled LDA (Ramage et al. 2009), and DMR (Mimno and McCallum 2008). The former two dedicate at least one topic to each metadata label, making them too computationally expensive to use on data sets with thousands of authors, such as the Blog and IMDb1M data sets (Section 5.1). In contrast to DiscLDA and Labeled LDA, DMR uses less topics by sharing them between labels. Mimno and McCallum (2008) showed that DMR outperformed AT on authorship attribution of multi-authored documents. Despite this, we decided to use AT, because we found in preliminary experiments that AT performs better than DMR on authorship attribution of single-authored texts. Such texts are the main focus of this article. Nonetheless, it is worth noting that Mimno and McCallum's experiments were performed on a data set of research papers where stopwords were filtered out. We do not discard stopwords in most experiments, because they are known to be indicators of authorship (Koppel, Schler, and Argamon 2009).[1]

A representative example of a generic downstream model is sLDA (Blei and McAuliffe 2007), which generates labels from each document's topic assignments via a generalized linear model. This model was extended by Zhu, Ahmed, and Xing (2009), who introduced MedLDA, where training is done in a way that maximizes the margin between labels, which is "arguably more suitable" for inferring document labels. Zhu

1 Following Salton (1981), we use the term **stopwords** to denote the most common words—we use his English stopword list from the SMART information retrieval system (Salton 1971), which is available from `www.lextek.com/manuals/onix/stopwords2.html`. Stopword lists typically include a set of non-content words, which is a superset of the function words in a given language. Although Koppel, Schler, and Argamon (2009) found that good performance can be obtained by relying only on function words, they also showed that the data-driven approach of relying on the most common words in a corpus yields superior performance in most cases. Either way, discarding stopwords is likely to yield poor results, as we show in Section 5.3.

and Xing (2010) further extended that work by introducing sCTRF, which combines sLDA with conditional random fields to accommodate arbitrary types of features. Zhu and Xing applied these models to the polarity inference task, and found that support vector regression outperformed sLDA and performed comparably to MedLDA (these three models used only unigrams), whereas sCTRF yielded the best performance by incorporating additional feature types (e.g., part-of-speech tags and a lexicon of positive and negative words). Based on these results, we decided to leave experiments with downstream models for future work, as it seems unlikely that we would obtain good results on the authorship attribution task without considering other feature types in addition to token unigrams (which is beyond the scope of this article).

3. Topic Models and Author Representations

This section introduces notation, provides a discussion of the meaning of the parameters used by the topic models, and describes the three topic models considered in this article (LDA, AT, and DADT), focusing on the author representations that they yield.

3.1 Notation and Preliminaries

We denote matrices and stacked vectors in uppercase boldface italics (e.g., $\boldsymbol{M}$), and vectors in lowercase boldface italics (e.g., $\boldsymbol{v}$). The element at the i-th row and j-th column of a matrix $\boldsymbol{M}$ is denoted m_{ij}, and vector elements are denoted in lowercase italics with a subscript index (e.g., v_i). Sets are denoted with calligraphic font (e.g., $\mathcal{S}$). In addition, Dir and Cat denote the Dirichlet and categorical distributions, respectively.

The values of the parameters that are given as input to the models are either determined by the corpus (Section 3.1.1) or configured when using the models (Sections 3.1.2 and 3.1.3). Table 1 shows the models' corpus-dependent and configurable parameters with their lengths and meanings (scalars have length 1). Corpus-dependent parameters are at the top, configurable document-related parameters are in the middle (for LDA

Table 1
Corpus-dependent and configurable parameters.

Symbol	Length	Meaning
A	1	Number of authors
D	1	Number of documents
V	1	Vocabulary size
N_d	1	Number of words in document d
$T^{(D)}$	1	Number of document topics
$\boldsymbol{\alpha}^{(D)}$	$T^{(D)}$	Document topic prior
$\boldsymbol{\beta}^{(D)}$	V	Word in document topic prior
$\delta^{(D)}$	1	Document words in document prior
$T^{(A)}$	1	Number of author topics
$\boldsymbol{\alpha}^{(A)}$	$T^{(A)}$	Author topic prior
$\boldsymbol{\beta}^{(A)}$	V	Word in author topic prior
$\boldsymbol{\eta}$	A	Author in corpus prior
$\delta^{(A)}$	1	Author words in document prior

and DADT), and configurable author-related parameters are at the bottom (for AT and DADT).

3.1.1 Corpus-Dependent Parameters. The following parameters depend on the corpus, and are thus considered to be observed:

- A: Number of authors. We use $a \in \{1, \ldots, A\}$ to denote an author identifier.
- D: Number of documents. We use $d \in \{1, \ldots, D\}$ to denote a document identifier.
- V: Vocabulary size. We use $v \in \{1, \ldots, V\}$ to denote a unique word identifier.
- N_d: Number of words in document d. We use $i \in \{1, \ldots, N_d\}$ to denote a word index in document d.
- $\boldsymbol{A}$: Document authors. This is a D-dimensional vector of vectors, where the d-th element $\boldsymbol{a}_d$ contains the authors of the d-th document. In cases where the corpus contains only single-authored texts, we use the scalar a_d to denote the author of the d-th document, since $\boldsymbol{a}_d$ is always of unit length.
- $\boldsymbol{W}$: Document words. This is a D-dimensional vector of vectors, where the d-th element $\boldsymbol{w}_d$ contains the words of the d-th document. The vector $\boldsymbol{w}_d$ is of length N_d, and $w_{di} \in \{1, \ldots, V\}$ is the i-th word in the d-th document.

3.1.2 Number of Topics. We make a distinction between *document* topics and *author* topics. In both cases, "topics" describe distributions over all the words in the vocabulary. The difference is that document topics are word distributions that arise from documents, while author topics are word distributions that characterize the authors. LDA uses only document topics, whereas AT uses only author topics. DADT, our hybrid model, uses both document topics and author topics.

All three models take the number of topics as a configurable parameter, denoted by $T^{(D)}$ for the number of document topics and by $T^{(A)}$ for the number of author topics. Although the models have other configurable parameters (introduced subsequently), we found that the number of topics has the largest impact on model performance because it controls the overall model complexity. For example, setting $T^{(D)} = 1$ in LDA means that all the words in all the documents are drawn from the same topic (i.e., a single distribution for all the words), whereas setting $T^{(D)} = 200$ gives LDA much more freedom to adapt to the corpus, as each word can be drawn from one of 200 distributions.

It is worth noting that techniques for determining the optimal number of topics have been suggested. For example, Teh et al. (2006) used hierarchical Dirichlet processes to learn the number of topics while inferring the LDA model. We did not experiment with such techniques as they tend to complicate model inference, and we found that using a constant number of topics yields good performance. Nonetheless, we note that utilizing such techniques may be a worthwhile future research direction, especially to determine the balance between document topics and author topics for DADT.

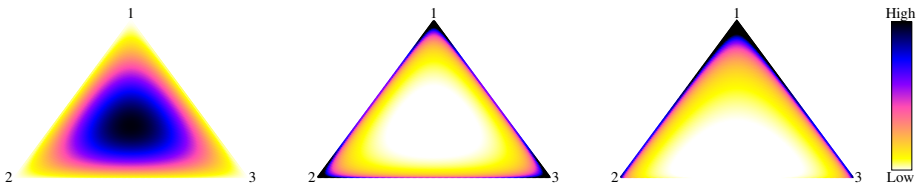
3.1.3 Distribution Priors. The following parameters are the priors of the Dirichlet and beta distributions used by the models. In contrast to the number of topics, which controls model complexity, the priors allow users of the models to specify their prior knowledge and beliefs about the data. In addition, the number of topics imposes a rigid constraint on the inferred model, whereas the effect of the priors on the model is expected to diminish as the amount of observed data increases (Equations (3), (7), and (11)). Indeed, we found in our experiments that varying prior values had a small effect on performance compared to varying the number of topics (Figure 8b, Section 5.3.4).

The priors are defined as follows (all vector elements and scalars are positive):

$\boldsymbol{\alpha}^{(D)}$: Document topic prior – a vector of length $T^{(D)}$.
$\boldsymbol{\beta}^{(D)}$: Prior for words in document topics – a vector of length V.
$\boldsymbol{\alpha}^{(A)}$: Author topic prior – a vector of length $T^{(A)}$.
$\boldsymbol{\beta}^{(A)}$: Prior for words in author topics – a vector of length V.
$\delta^{(D)}$: Document words in document prior.
$\delta^{(A)}$: Author words in document prior.
$\boldsymbol{\eta}$: Author in corpus prior – a vector of length A.

The support of a K-dimensional Dirichlet distribution Dir($\boldsymbol{\alpha}$) is the set of K-dimensional vectors with elements in the range $[0, 1]$ whose sum is 1 (the Dirichlet distribution is a multivariate generalization of the beta distribution). Hence, each draw from the Dirichlet distribution can be seen as defining the parameters of a categorical distribution. This is illustrated by Figure 1, which shows the Dirichlet distribution density in the three-dimensional case for three different prior vectors $\boldsymbol{\alpha}$ (the density is triangular because the drawn vector elements have to sum to 1—each corner of the triangle corresponds to a dimension of the distribution, denoted 1, 2, and 3 in Figure 1). When the prior vector is symmetric (i.e., all its elements have the same value), the density is also symmetric (Figures 1a and 1b). Symmetric priors with element values that are greater than 1 yield densities that are concentrated in the middle of the triangle, meaning that categorical vectors with relatively uniform values are likely to be drawn (Figure 1a). On the other hand, symmetric priors with element values that are less than 1 yield sparse densities with high values in the corners of the triangle, meaning that the categorical vectors are likely to have one element whose value is greater than the other elements (Figure 1b). Finally, when the prior is asymmetric, vectors that give higher probabilities to the elements with higher prior values are likely to be drawn (Figure 1c).

The document and author topic priors ($\boldsymbol{\alpha}^{(D)}$ and $\boldsymbol{\alpha}^{(A)}$, respectively) encode our beliefs about the document and author topic distributions, respectively. They are often set to be symmetric, because we have no reason to favor one topic over the other before we have seen the data (Steyvers and Griffiths 2007). Wallach, Mimno, and McCallum (2009) argue that using asymmetric priors in LDA is beneficial, and suggest a method that learns such priors as part of model inference (by placing another prior on the $\boldsymbol{\alpha}^{(D)}$ prior). We implemented Wallach, Mimno and McCallum's method for all the models we considered, but found that it did not improve authorship attribution accuracy in preliminary experiments. Thus, in all our experiments we set 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, yielding relatively sparse topic distributions, since we expect each document and author to be sufficiently



(a) $\boldsymbol{\alpha}$ = (2, 2, 2) (b) $\boldsymbol{\alpha}$ = (0.5, 0.5, 0.5) (c) $\boldsymbol{\alpha}$ = (0.9, 0.5, 0.5)

Figure 1
Three-dimensional Dirichlet probability density, given three prior vectors.

represented by only a few topics. This choice follows the recommendations from LingPipe's documentation (`alias-i.com/lingpipe`), which are based on empirical evidence from several corpora.

The priors for words in document and author topics ($\boldsymbol{\beta}^{(D)}$ and $\boldsymbol{\beta}^{(A)}$, respectively) encode our beliefs about the word distributions. As for the topic distribution priors, symmetric priors are often used, with a default value of 0.01 for all the vector elements (yielding sparse word distributions, as indicated earlier), meaning that each topic is expected to assign high probabilities to only a few top words (Steyvers and Griffiths 2007). In contrast to the topic distribution priors, Wallach, Mimno, and McCallum (2009) found in their experiments on LDA that using an asymmetric $\boldsymbol{\beta}^{(D)}$ was of no benefit. This is because using an asymmetric $\boldsymbol{\beta}^{(D)}$ means that we encode a prior preference for a certain word to appear in all topics (e.g., a word represented by corner 1 in Figure 1c). For the same reason, using a symmetric $\boldsymbol{\beta}^{(A)}$ is a sensible choice for AT. In contrast to LDA and AT, our DADT model distinguishes between document words and author words, and thus uses both $\boldsymbol{\beta}^{(D)}$ and $\boldsymbol{\beta}^{(A)}$ as priors. This allows us to encode our prior knowledge that stopword use is indicative of authorship. Thus, for DADT we set $\beta_v^{(D)} = 0.01 - \epsilon$ and $\beta_v^{(A)} = 0.01 + \epsilon$ for all v, where v is a stopword (ϵ can be set to zero to obtain symmetric priors).

DADT's $\delta^{(D)}$ and $\delta^{(A)}$ priors encode our prior belief about the balance between document words and author words in a given document. Document words (drawn from document topics) are expected to be representative of the documents in the corpus, whereas author words (drawn from author topics) characterize the authors in the corpus. For example, if we asked two different authors to write a report about LDA, both reports are likely to contain content words like *Dirichlet*, *topic*, and *prior*, but the frequencies of non-content words (i.e., function words and other indicators of authorship style) are likely to vary across the reports. In this case, the content words are expected to be allocated to document topics, and the non-content words whose usage varies across authors would be allocated to author topics. In cases where the authors write about different issues, DADT may allocate some content words to author topics (i.e., the meaning of DADT's topics is expected to be corpus-specific). According to DADT's definition (Section 3.4.1), which uses the beta distribution, the prior expected value of the portion of each document that is composed of author words is

$$\frac{\delta^{(A)}}{\delta^{(A)} + \delta^{(D)}} \tag{1}$$

with a variance of

$$\frac{\delta^{(A)}\delta^{(D)}}{\left(\delta^{(A)} + \delta^{(D)}\right)^2 \left(\delta^{(A)} + \delta^{(D)} + 1\right)} \tag{2}$$

In our experiments, we chose values for $\delta^{(D)}$ and $\delta^{(A)}$ by deciding on the expected value and variance, and solving these equations for $\delta^{(D)}$ and $\delta^{(A)}$.

Finally, DADT's $\boldsymbol{\eta}$ prior determines the prior belief about an author having written a document (without looking at the actual words in the document). This prior is only used on documents with unobserved authors (i.e., when attributing authors to anonymous texts). Because we have no reason to favor one author over the other, we use a uniform prior, setting $\eta_a = 1$ for each author a.

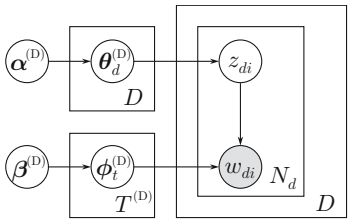


Figure 2
Latent Dirichlet allocation (LDA).

3.2 LDA

3.2.1 Model Definition. LDA was originally defined by Blei, Ng, and Jordan (2003). Here we describe Griffiths and Steyvers's (2004) extended version. The idea behind LDA is that each document in a corpus is described by a distribution over topics, and each word in the document is drawn from its topic's word distribution. Figure 2 presents LDA in plate notation, where observed variables are in shaded circles, unobserved variables are in unshaded circles, and each box represents repeated sampling, with the number of repetitions at the bottom-right corner. Formally, the generative process is: (1) for each topic t, draw a word distribution $\boldsymbol{\phi}_t^{(D)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(D)}\right)$; (2) for each document d, draw a topic distribution $\boldsymbol{\theta}_d^{(D)} \sim \text{Dir}\left(\boldsymbol{\alpha}^{(D)}\right)$; and (3) for each word index i in each document d, draw a topic $z_{di} \sim \text{Cat}\left(\boldsymbol{\theta}_d^{(D)}\right)$, and the word $w_{di} \sim \text{Cat}\left(\boldsymbol{\phi}_{z_{di}}^{(D)}\right)$.

3.2.2 Model Inference. Topic models are commonly inferred using either collapsed Gibbs sampling (Griffiths and Steyvers 2004; Rosen-Zvi et al. 2004) or methods based on variational inference (Blei, Ng, and Jordan 2003). We use collapsed Gibbs sampling to infer all models due to its efficiency and ease of implementation. This involves repeatedly sampling from the conditional distribution of the latent parameters, which is obtained analytically by marginalizing over the topic and word distributions, and using the properties of conjugate priors. This conditional distribution is given in Equation (3) (Griffiths and Steyvers 2004; Steyvers and Griffiths 2007):

$$p\left(z_{di} = t | \boldsymbol{W}, \boldsymbol{Z}_{-di}; \boldsymbol{\alpha}^{(D)}, \boldsymbol{\beta}^{(D)}\right) \propto \frac{\alpha_t^{(D)} + c_{dt}^{(DT)}}{\sum_{t'=1}^{T^{(D)}} \left(\alpha_{t'}^{(D)} + c_{dt'}^{(DT)}\right)} \frac{\beta_{w_{di}}^{(D)} + c_{tw_{di}}^{(DTV)}}{\sum_{v=1}^{V} \left(\beta_v^{(D)} + c_{tv}^{(DTV)}\right)} \tag{3}$$

where $\boldsymbol{W}$ is the corpus; $\boldsymbol{Z}_{-di}$ contains all the topic assignments, excluding the assignment for the i-th word of the d-th document; $c_{dt}^{(DT)}$ is the count of topic t in document d; and $c_{tw_{di}}^{(DTV)}$ is the count of word w_{di} in document topic t. Here, these counts exclude the di-th topic assignment (i.e., z_{di}).

Commonly, several Gibbs sampling chains are run, and several samples are retained from each chain after a burn-in period, which allows the chain to reach its stationary distribution. For each sample, the topic distributions and the word distributions are estimated using their expected values, given the topic assignments $\boldsymbol{Z}$. These expected values are given in Equations (4) and (5):

$$\mathbb{E}[\theta_{dt}^{(D)} | \boldsymbol{Z}] = \frac{\alpha_t^{(D)} + c_{dt}^{(DT)}}{\sum_{t'=1}^{T^{(D)}} \left(\alpha_{t'}^{(D)} + c_{dt'}^{(DT)}\right)} \tag{4}$$

$$\mathbb{E}[\phi_{tv}^{(D)}|\mathbf{Z}] = \frac{\beta_v^{(D)} + c_{tv}^{(DTV)}}{\sum_{v'=1}^{V}\left(\beta_{v'}^{(D)} + c_{tv'}^{(DTV)}\right)} \tag{5}$$

where in this case the counts are over the full topic and author assignments. The two equations take a similar form due to the fact that the Dirichlet distribution is the conjugate prior of the categorical distribution (Griffiths and Steyvers 2004). Note that these values cannot be averaged across samples due to the exchangeability of the topics (Steyvers and Griffiths 2007) (e.g., topic 1 in one sample is not necessarily the same as topic 1 in another sample).

The examined authorship attribution problem follows a supervised classification setup, where training texts with known candidate authors are given in advance. Test texts are classified one by one, and the goal is to attribute each test text to one of the candidate authors. As the word distributions of the LDA model inferred in the training phase are unlikely to change much due to the addition of a single test document, in the classification phase we consider each topic's word distribution to be observed, setting it to its expected value according to Equation (5). This yields the following sampling equation for a given test text $\tilde{\boldsymbol{w}}$ ($\tilde{\boldsymbol{w}}$ is a word vector of length $\tilde{N}$):

$$p\left(\tilde{z}_i = t|\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}_{-i}; \boldsymbol{\Phi}^{(D)}, \boldsymbol{\alpha}^{(D)}\right) \propto \frac{\alpha_t^{(D)} + \tilde{c}_t^{(DT)}}{\sum_{t'=1}^{T^{(D)}}\left(\alpha_{t'}^{(D)} + \tilde{c}_{t'}^{(DT)}\right)} \phi_{t\tilde{w}_i}^{(D)} \tag{6}$$

where $\tilde{z}_i$ is the topic assignment for the i-th word in $\tilde{\boldsymbol{w}}$, $\tilde{\boldsymbol{z}}_{-i}$ contains all of $\tilde{\boldsymbol{w}}$'s topic assignments except for the i-th assignment, and $\tilde{c}_t^{(DT)}$ is the count of words assigned to topic t, excluding the i-th assignment.

As done in the training phase, we set the test text's topic distribution $\tilde{\boldsymbol{\theta}}^{(D)}$ to its expected value according to Equation (4), where $c_{dt}^{(DT)}$ is replaced with $\tilde{c}_t^{(DT)}$ (which now contains the counts over the full vector of topic assignments $\tilde{\boldsymbol{z}}$). Note that because we assume that the $\boldsymbol{\phi}_t^{(D)}$ values are observed in the classification phase, the topics are *not* exchangeable. This means that we can average the $\mathbb{E}[\tilde{\boldsymbol{\theta}}_t^{(D)}|\tilde{\boldsymbol{z}}]$ values across test samples obtained from the same sampling chain.

3.2.3 Author Representations. LDA does not directly model authors, but it can still be used to obtain valuable information about them. The output of LDA consists of distributions over topics $\boldsymbol{\theta}_d^{(D)}$ for each document d. As the number of topics $T^{(D)}$ is commonly much smaller than the size of the vocabulary V, these topical representations form a lower-dimensional representation of the corpus. The LDA-based author representation we consider in this article is **LDA-M** (LDA with multiple documents per author), where each author a is represented as the set of distributions over topics of their documents, namely, the set $\{\boldsymbol{\theta}_d^{(D)}|a_d = a\}$, where a_d is the author of document d. An alternative approach is **LDA-S** (LDA with a single document per author), where each author's documents are concatenated into a single document in a preprocessing step, LDA is run on the concatenated documents, and each author is represented by a single distribution over topics (the distribution of the concatenated document).

An advantage of LDA-S over LDA-M is that LDA-S yields a much more compact author representation than LDA-M, especially for authors who wrote many documents. However, this compactness may come at the price of accuracy, as markers that may be present only in a few short documents by one author may lose their prominence

if these documents are concatenated with longer documents. It is worth noting that concatenating each author's documents into one document has been named the **profile-based** approach in previous authorship attribution studies, in contrast to the **instance-based** approach, where each document is considered separately (Stamatatos 2009).

A limitation of these representations is that they apply only to corpora of single-authored documents, and there is no straightforward way of extending them to consider multi-authored documents. This limitation is addressed by AT, which we present in the next section. Note that when analyzing single-authored documents, the author representations yielded by AT are equivalent to LDA-S's representations. Therefore, we do not report results obtained with LDA-S. Nonetheless, practitioners may find it easier to use LDA-S than AT due to the relative prevalence of LDA implementations (in fact, our initial modeling approach was LDA-S for exactly this reason).

3.3 AT

3.3.1 Model Definition. AT was introduced by Rosen-Zvi et al. (2004) to model author interests in corpora of multi-authored texts (e.g., research papers). The main idea behind AT is that each document is generated from the topic distributions of its observed authors, rather than from a document-specific topic distribution. Figure 3 presents AT in plate notation. Formally, the generative process is: (1) for each topic t, draw a word distribution $\boldsymbol{\phi}_t^{(A)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(A)}\right)$; (2) for each author a, draw a topic distribution $\boldsymbol{\theta}_a^{(A)} \sim \text{Dir}\left(\boldsymbol{\alpha}^{(A)}\right)$; and (3) for each word index i in each document d, draw an author x_{di} uniformly from the document's set of authors $\boldsymbol{a}_d$, a topic $z_{di} \sim \text{Cat}\left(\boldsymbol{\theta}_{x_{di}}^{(A)}\right)$, and the word $w_{di} \sim \text{Cat}\left(\boldsymbol{\phi}_{z_{di}}^{(A)}\right)$.

3.3.2 Model Inference. As for LDA, we use collapsed Gibbs sampling to infer AT. This involves repeatedly sampling from Equation (7) (Rosen-Zvi et al. 2004, 2010):

$$p\left(\begin{matrix} x_{di} = a, \\ z_{di} = t \end{matrix} \middle| \begin{matrix} \boldsymbol{A}, \boldsymbol{W}, \boldsymbol{X}_{-di}, \boldsymbol{Z}_{-di}; \\ \boldsymbol{\alpha}^{(A)}, \boldsymbol{\beta}^{(A)} \end{matrix}\right) \propto \frac{\alpha_t^{(A)} + c_{at}^{(AT)}}{\sum_{t'=1}^{T^{(A)}} \left(\alpha_{t'}^{(A)} + c_{at'}^{(AT)}\right)} \frac{\beta_{w_{di}}^{(A)} + c_{tw_{di}}^{(ATV)}}{\sum_{v=1}^{V} \left(\beta_v^{(A)} + c_{tv}^{(ATV)}\right)} \tag{7}$$

where $\boldsymbol{X}_{-di}$ and $\boldsymbol{Z}_{-di}$ are all the author and topic assignments, respectively, excluding the assignment for the i-th word of the d-th document; $c_{at}^{(AT)}$ is the count of topic t assignments to author a; and $c_{tw_{di}}^{(ATV)}$ is the count of word w_{di} in author topic t. Here, all the counts exclude the di-th assignments (i.e., x_{di} and z_{di}). We sample x_{di} and z_{di}

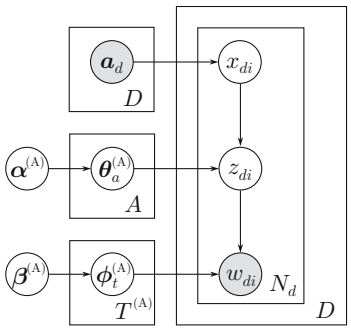


Figure 3
The Author-Topic (AT) model.

jointly because this yields faster convergence than separate sampling (Rosen-Zvi et al. 2010).

Similarly to LDA, we estimate the topic and word distributions using their expected values given the author assignments $\boldsymbol{X}$ and the topic assignments $\boldsymbol{Z}$:

$$\mathbb{E}[\theta_{at}^{(A)}|\boldsymbol{X},\boldsymbol{Z}] = \frac{\alpha_t^{(A)} + c_{at}^{(AT)}}{\sum_{t'=1}^{T^{(A)}} \left(\alpha_{t'}^{(A)} + c_{at'}^{(AT)}\right)} \tag{8}$$

$$\mathbb{E}[\phi_{tv}^{(A)}|\boldsymbol{Z}] = \frac{\beta_v^{(A)} + c_{tv}^{(ATV)}}{\sum_{v'=1}^{V} \left(\beta_{v'}^{(A)} + c_{tv'}^{(ATV)}\right)} \tag{9}$$

where in this case the counts are over the full author and topic assignments.

In the classification phase, we do not know the author $\tilde{a}$ of the test text $\tilde{\boldsymbol{w}}$ (we assume that test texts are single-authored). If we did, no sampling would be required to obtain $\tilde{a}$'s topic distribution because it is already inferred in the training phase (Equation (8)). Hence, we assume that $\tilde{a}$ is a "new," previously unknown author, and utilize Gibbs sampling to infer this author's topic distribution $\tilde{\boldsymbol{\theta}}^{(A)}$ by repeatedly sampling from Equation (10) (as for LDA, the word distributions are assumed to be observed and set to their expected values according to Equation (9)):

$$p\left(\tilde{z}_i = t|\tilde{\boldsymbol{w}},\tilde{\boldsymbol{z}}_{-i};\boldsymbol{\Phi}^{(A)},\boldsymbol{\alpha}^{(A)}\right) \propto \frac{\alpha_t^{(A)} + \tilde{c}_t^{(AT)}}{\sum_{t'=1}^{T^{(A)}} \left(\alpha_{t'}^{(A)} + \tilde{c}_{t'}^{(AT)}\right)} \phi_{t\tilde{w}_i}^{(A)} \tag{10}$$

where $\tilde{z}_i$ is the topic assignment for the i-th word in $\tilde{\boldsymbol{w}}$, $\tilde{\boldsymbol{z}}_{-i}$ contains all of $\tilde{\boldsymbol{w}}$'s topic assignments except for the i-th assignment, and $\tilde{c}_t^{(AT)}$ is the count of topic t assignments to author $\tilde{a}$ (excluding the i-th assignment). Similarly to LDA, we then set $\tilde{\boldsymbol{\theta}}^{(A)}$ to its expected value according to Equation (8), where $c_{at}^{(AT)}$ is replaced with $\tilde{c}_t^{(AT)}$ over the full assignment vector $\tilde{\boldsymbol{z}}$.

3.3.3 Author Representations. AT naturally yields author representations in the form of distributions over topics. That is, each author a is represented as a distribution over topics $\boldsymbol{\theta}_a^{(A)}$. However, AT is limited because all the documents by the same authors are generated in an identical manner (Section 3.3.1). To address this limitation, Rosen-Zvi et al. (2010) introduced "fictitious" authors, adding a unique "author" to each document. This allows AT to adapt itself to each document without changing the model specification. Therefore, we consider the two following variants: (1) **AT**: "Pure" AT, without fictitious authors; and (2) **AT-FA**: AT, when run with the additional preprocessing step of adding a fictitious author to each document.

3.4 DADT

3.4.1 Model Definition. Our DADT model can be seen as a combination of LDA and AT, which is meant to address the weaknesses of both models while retaining their strengths. The main idea behind DADT is that words are generated from two disjoint sets of topics: document topics and author topics. Words generated from document topics follow the same generation process as in LDA, whereas words generated from

author topics are generated in an AT-like fashion. This approach has the potential benefit of separating "document" words from "author" words. That is, words whose use varies across documents are expected to be found in document topics, whereas words whose use varies between authors are expected to be assigned to author topics. Figure 4 presents the graphical representation of the model, where the document-dependent parameters appear on the left-hand side, and the author-dependent parameters appear on the right-hand side. Formally, the generative process is as follows (we mark each step as coming from either **L**DA or **A**T, or as new in **D**ADT).

Corpus level:

L. For each document topic t, draw a word distribution $\boldsymbol{\phi}_t^{(D)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(D)}\right)$.
A. For each author topic t, draw a word distribution $\boldsymbol{\phi}_t^{(A)} \sim \text{Dir}\left(\boldsymbol{\beta}^{(A)}\right)$.
A. For each author a, draw an author topic distribution $\boldsymbol{\theta}_a^{(A)} \sim \text{Dir}\left(\boldsymbol{\alpha}^{(A)}\right)$.
D. Draw an author distribution $\boldsymbol{\chi} \sim \text{Dir}\left(\boldsymbol{\eta}\right)$.

Document level. For each document d:

L. Draw d's document topic distribution $\boldsymbol{\theta}_d^{(D)} \sim \text{Dir}\left(\boldsymbol{\alpha}^{(D)}\right)$.
D. Draw d's author set $\boldsymbol{a}_d$ by repeatedly sampling without replacement from $\text{Cat}\left(\boldsymbol{\chi}\right)$.
D. Draw d's author/document topic ratio $\pi_d \sim \text{Beta}\left(\delta^{(A)}, \delta^{(D)}\right)$.

Word level. For each word index $i \in \{1, \ldots, N_d\}$:

D. Draw the author/document topic indicator $y_{di} \sim \text{Bernoulli}(\pi_d)$.
L. If $y_{di} = 0$, use document topics: draw a topic $z_{di} \sim \text{Cat}\left(\boldsymbol{\theta}_d^{(D)}\right)$, and the word $w_{di} \sim \text{Cat}\left(\boldsymbol{\phi}_{z_{di}}^{(D)}\right)$.
A. If $y_{di} = 1$, use author topics: Draw an author x_{di} uniformly from $\boldsymbol{a}_d$, a topic $z_{di} \sim \text{Cat}\left(\boldsymbol{\theta}_{x_{di}}^{(A)}\right)$, and the word $w_{di} \sim \text{Cat}\left(\boldsymbol{\phi}_{z_{di}}^{(A)}\right)$.

It is worth noting that drawing the document's author set can also be modeled as sampling from Wallenius's noncentral hypergeometric distribution (Fog 2008) with a weight vector $\boldsymbol{\chi}$ and a parameter vector whose elements are all equal to 1. In this article, we consider only situations where $\boldsymbol{a}_d$ is observed when the model is inferred. When handling documents with unknown authors in our authorship attribution experiments, we assume that all anonymous texts are single-authored.

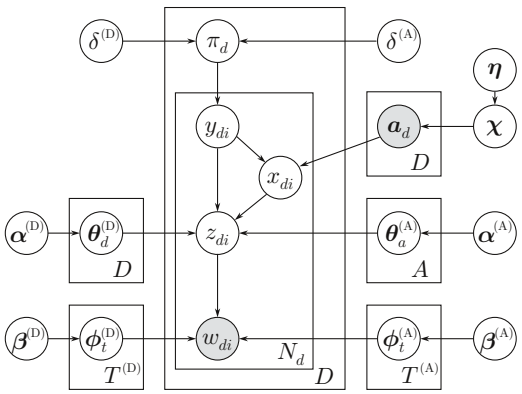


Figure 4
The Disjoint Author-Document Topic (DADT) model.

3.4.2 Model Inference. We infer DADT using collapsed Gibbs sampling, as done for LDA and AT. This involves repeatedly sampling from the following conditional distribution of the latent parameters:

$$p\left(\begin{matrix} x_{di}=a, \\ y_{di}=y, z_{di}=t \end{matrix} \middle| \begin{matrix} \boldsymbol{A},\boldsymbol{W},\boldsymbol{X}_{-di},\boldsymbol{Y}_{-di},\boldsymbol{Z}_{-di}; \\ \boldsymbol{\alpha}^{(D)},\boldsymbol{\beta}^{(D)},\delta^{(D)},\boldsymbol{\alpha}^{(A)},\boldsymbol{\beta}^{(A)},\delta^{(A)} \end{matrix}\right) \propto \tag{11}$$

$$\begin{cases} \left(\delta^{(D)} + c_d^{(DD)}\right) \frac{\alpha_t^{(D)} + c_{dt}^{(DT)}}{\sum_{t'=1}^{T^{(D)}} \left(\alpha_{t'}^{(D)} + c_{dt'}^{(DT)}\right)} \frac{\beta_{w_{di}}^{(D)} + c_{tw_{di}}^{(DTV)}}{\sum_{v=1}^{V} \left(\beta_v^{(D)} + c_{tv}^{(DTV)}\right)} & \text{if } y = 0 \\ \left(\delta^{(A)} + c_d^{(DA)}\right) \frac{\alpha_t^{(A)} + c_{at}^{(AT)}}{\sum_{t'=1}^{T^{(A)}} \left(\alpha_{t'}^{(A)} + c_{at'}^{(AT)}\right)} \frac{\beta_{w_{di}}^{(A)} + c_{tw_{di}}^{(ATV)}}{\sum_{v=1}^{V} \left(\beta_v^{(A)} + c_{tv}^{(ATV)}\right)} & \text{if } y = 1 \end{cases}$$

where $\boldsymbol{Y}_{-di}$ contains the topic indicators, excluding the di-th value; and $c_d^{(DD)}$ and $c_d^{(DA)}$ are the counts of words assigned to document or author topics in document d, respectively. The other variables are defined as for LDA and AT. Here, all the counts exclude the di-th assignments (i.e., x_{di}, y_{di}, and z_{di}).

The building blocks of our DADT model are clearly visible in Equation (11). LDA's Equation (3) is contained in the $y = 0$ case, where the word is drawn from document topics, and AT's Equation (7) is contained in the $y = 1$ case, where the word is drawn from author topics. However, Equation (11) also demonstrates the main difference between DADT and its building blocks, as DADT considers *both* documents and authors during the inference process by assigning each word to either a document topic or an author topic, where document topics and author topics come from disjoint sets.

As for LDA and AT, we ran several sampling chains in our experiments, retaining samples from each chain after a burn-in period (a sample consists of $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$). For each sample, the topic and word distributions are estimated using their expected values given the latent variable assignments. The expected values for the topic and word distributions are the same as for LDA and AT, and the expected values for the author/document ratio and the corpus author distribution are:

$$\mathbb{E}[\pi_d|\boldsymbol{Y}] = \frac{\delta^{(A)} + c_d^{(DA)}}{\delta^{(D)} + \delta^{(A)} + N_d} \tag{12}$$

$$\mathbb{E}[\chi_a|\boldsymbol{W}] = \frac{\eta_a + c_a^{(AD)}}{\sum_{a'=1}^{A} \left(\eta_{a'} + c_{a'}^{(AD)}\right)} \tag{13}$$

where the counts are now over the full assignments $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$. As in LDA and AT, these equations were straightforward to obtain, because the Dirichlet distribution is the conjugate prior of the categorical distribution and the beta distribution is the conjugate prior of the Bernoulli distribution. It is worth noting that because we assume that the documents' authors are observed during model inference, the expected value of each element of the corpus distribution over authors χ_a does not vary across samples, as it only depends on the prior η_a and on author a's count of documents in the corpus $c_a^{(AD)}$.

In the classification phase, we do not know the author $\tilde{a}$ of the test text $\tilde{\boldsymbol{w}}$. As for AT, we assume that $\tilde{a}$ is a previously unknown author. We set the word distributions to their expected values from the training phase, and infer $\tilde{a}$'s author topic distribution $\tilde{\boldsymbol{\theta}}^{(A)}$

together with the test text's document topic distribution $\tilde{\boldsymbol{\theta}}^{(D)}$ and author/document topic ratio $\tilde{\pi}$ by repeatedly sampling from

$$p\left(\tilde{y}_i = y, \tilde{z}_i = t | \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{y}}_{-i}, \tilde{\boldsymbol{z}}_{-i}; \boldsymbol{\Phi}^{(D)}, \boldsymbol{\Phi}^{(A)}, \boldsymbol{\alpha}^{(D)}, \boldsymbol{\alpha}^{(A)}, \delta^{(D)}, \delta^{(A)}\right) \propto \tag{14}$$

$$\begin{cases} \left(\delta^{(D)} + \tilde{c}^{(DD)}\right) \frac{\alpha_t^{(D)} + \tilde{c}_t^{(DT)}}{\sum_{t'=1}^{T^{(D)}} \left(\alpha_{t'}^{(D)} + \tilde{c}_{t'}^{(DT)}\right)} \phi_{t\tilde{w}_i}^{(D)} & \text{if } y = 0 \\ \left(\delta^{(A)} + \tilde{c}^{(DA)}\right) \frac{\alpha_t^{(A)} + \tilde{c}_t^{(AT)}}{\sum_{t'=1}^{T^{(A)}} \left(\alpha_{t'}^{(A)} + \tilde{c}_{t'}^{(AT)}\right)} \phi_{t\tilde{w}_i}^{(A)} & \text{if } y = 1 \end{cases}$$

where $\tilde{y}_i$ is the topic indicator for the i-th word, $\tilde{\boldsymbol{y}}_{-i}$ contains all of $\tilde{\boldsymbol{w}}$'s topic indicators except for the i-th indicator, and $\tilde{c}^{(DD)}$ and $\tilde{c}^{(DA)}$ are the counts of words assigned to document and author topics, respectively, excluding the i-th assignment (the other variables are defined as in Equations (6) and (10)). The expected values of $\tilde{\boldsymbol{\theta}}^{(D)}$ and $\tilde{\boldsymbol{\theta}}^{(A)}$ are the same as for LDA and AT, respectively. The expected value of $\tilde{\pi}$ is obtained by replacing $c_d^{(DA)}$ and N_d with $\tilde{c}^{(DA)}$ and $\tilde{N}$ in Equation (12) (where $\tilde{c}^{(DA)}$ now contains the counts over the full vector of indicators $\tilde{\boldsymbol{y}}$).

3.4.3 Author Representations and Comparison to LDA and AT. DADT can be seen as a generalization of LDA and AT—setting DADT's number of author topics $T^{(A)}$ to zero yields a model that is equivalent to LDA, and setting the number of document topics $T^{(D)}$ to zero yields a model that is equivalent to AT. An advantage of DADT over LDA and AT is that both documents and authors are accounted for in the model's definition, and are represented via distributions over document and author topics, respectively. Hence, preprocessing steps such as concatenating each author's documents or adding fictitious authors—as done in LDA-S and AT-FA to obtain author and document representations, respectively—are unnecessary.

Of the LDA and AT variants presented in Sections 3.2.3 and 3.3.3, DADT might seem most similar to AT-FA. However, there are several key differences between DADT and AT-FA.

First, in DADT, *author topics are disjoint from document topics*, with different priors for each topic set. Thus, the number of author topics $T^{(A)}$ can be different from the number of document topics $T^{(D)}$, which enables us to vary the number of author and document topics according to the number of authors and documents in the corpus. For example, in the judgment data set (Section 5.1.1), which includes only a few authors that wrote many long documents, we expect that small values of $T^{(A)}$ compared to $T^{(D)}$ would suffice to get good author representations. By contrast, modeling the 19,320 authors of the Blog data set (Section 5.1.5) is expected to require many more author topics. On such large data sets, using more than a few hundred topics may become too computationally expensive, because adding topics increases model complexity and thus adds to the runtime of the inference algorithm. Hence, being able to specify the balance between document and author topics in such cases is beneficial (Section 5.4).

Second, DADT *places different priors on the word distributions* for author topics and document topics ($\boldsymbol{\beta}^{(A)}$ and $\boldsymbol{\beta}^{(D)}$, respectively). We know from previous work that stopwords are strong indicators of authorship (Koppel, Schler, and Argamon 2009). Our model allows us to encode this prior knowledge by giving elements that correspond to stopwords in $\boldsymbol{\beta}^{(A)}$ higher weights than such elements in $\boldsymbol{\beta}^{(D)}$. We found that this property of DADT has practical benefits, as it improved the accuracy of DADT-based authorship attribution methods in our experiments (Section 5).

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 show that this has practical benefits in our authorship attribution experiments (Section 5): Specifying a prior belief that on average about 80% of each document is composed of author words can yield better results than using AT's fictitious author approach that 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 authorship attribution. In addition, this enables the use of DADT in a semi-supervised setup by training on documents with unknown authors—an extension that is left for future work.

4. Authorship Attribution Methods

This section introduces the authorship attribution methods considered in this article. In Section 4.1, we discuss our baseline method (SVM trained on tokens), and Sections 4.2, 4.3, 4.4, and 4.5 introduce methods based on LDA, AT, AT-FA, and DADT, respectively. These methods are summarized in Table 2.

We consider two approaches to using topic models in authorship attribution: dimensionality reduction and probabilistic.

Under the **dimensionality reduction** approach, the original documents are converted to topic distributions, and the topic distributions are used as input to a classifier. Generally, this approach makes it possible to use classifiers that are too computationally expensive to use with a large feature set, e.g., Webb, Boughton and Wang's (2005) AODE classifier, whose time complexity is quadratic in the number of features. We use the reduced document representations as input to SVM, and compare their performance with the performance obtained with SVM trained directly on tokens (denoted *Token SVM*). This allows us to roughly gauge how much information is lost by converting texts from token representations to topic representations. However, this approach ignores the probabilistic nature of the underlying topic model, and thus does not fully test the utility of the author representations yielded by the model—these are better tested by the next approach.

Table 2
Summary of authorship attribution methods.

Method	Description
Token SVM	Baseline: SVM trained on token frequencies
LDA-SVM	SVM trained on LDA document topic distributions
AT-SVM	SVM trained on AT author topic distributions
AT-P	Probabilistic attribution with AT
AT-FA-SVM	SVM trained on AT-FA author topic distributions (real and fictitious)
AT-FA-P1	Probabilistic attribution with AT-FA (classification *without* fictitious authors)
AT-FA-P2	Probabilistic attribution with AT-FA (classification *with* fictitious authors)
DADT-SVM	SVM trained on DADT document and author topic distributions
DADT-P	Probabilistic attribution with DADT

In contrast to dimensionality reduction methods, **probabilistic** methods utilize the underlying model's definitions directly to estimate the probability that a given author wrote a given test text. These methods require the model to be aware of authors, which means that LDA cannot be used in this case. We expect this approach to outperform the dimensionality reduction approach because the probabilistic approach considers the structure of the topic model.

An alternative approach that we considered uses a distance measure (e.g., Hellinger distance) to find the author whose topic distributions are closest to the distributions inferred from the test text. We do not describe distance-based methods in this article because we found that they yield poor results in most cases (Seroussi 2012), probably because they do not fully consider the underlying structure of the topic model.

4.1 Baseline: Token SVM

Our baseline method is SVM trained on token frequency features (i.e., token counts divided by the total number of tokens in the document). This method is known to yield state-of-the-art authorship attribution performance on this feature set; that is, when comparing methods without any further feature engineering, Token SVM is expected to yield good performance with minimal tuning (Koppel, Schler, and Argamon 2009). We use the one-versus-all setup to handle non-binary authorship attribution scenarios. This setup scales linearly in the number of authors and was shown to be at least as effective as other multi-class SVM approaches in many cases (Rifkin and Klautau 2004).

It is worth noting that unlike the topic models, the Token SVM baseline is trained with the goal of maximizing the authorship attribution accuracy, which may give Token SVM an advantage over topic-based methods. Further, as a discriminative classification approach, SVM may yield better performance than probabilistic topic-based methods, which are generative classifiers (Ng and Jordan 2001). However, as demonstrated by Ng and Jordan's comparison of discriminative and generative classifiers, this better performance may only be obtained in the presence of "enough" training data (just how much data is "enough" depends on the data set).

4.2 Methods Based on LDA

4.2.1 Dimensionality Reduction: LDA-SVM. Using LDA for dimensionality reduction is relatively straightforward—all it entails is converting the training and test texts to topic distributions as described in Section 3.2.2, and using these topic distributions as classifier features. Because we use SVM, it is possible to directly compare the results obtained with the LDA-SVM method to the baseline results obtained by running SVM trained directly on token frequencies.

This LDA-SVM approach was utilized by Blei, Ng, and Jordan (2003) to demonstrate the dimensionality reduction capabilities of LDA on the task of classifying articles according to a set of predefined categories. To the best of our knowledge, only Rajkumar et al. (2009) have previously applied LDA-SVM to authorship attribution—they published preliminary results obtained by running LDA-SVM, but did not compare their results to a Token SVM baseline. In Section 5, we present the results of more extensive experiments on the applicability of this approach to authorship attribution.

4.3 Methods Based on AT

4.3.1 Dimensionality Reduction: AT-SVM. We cannot use AT to obtain *document* topic distributions, because AT only infers *author* topic distributions (Section 3.3). Hence, we train the SVM component on the author topic representations (each document is

converted to its author topic distribution). For each test text, we assume that it was written by a previously unknown author, infer this author's topic distribution $\tilde{\boldsymbol{\theta}}^{(A)}$ (Section 3.3.2), and classify this distribution. This may be seen as very radical dimensionality reduction, because each author's entire set of training documents is reduced to a single author topic distribution.

4.3.2 Probabilistic: AT-P. For each author a, AT-P calculates the probability of the test text words given the AT model inferred from the training texts, under the assumption that the test text was written by a. It returns the author for whom this probability is the highest:

$$\underset{a\in\{1,\ldots,A\}}{\arg\max}\; p\left(\tilde{\boldsymbol{w}}|\tilde{a}=a,\boldsymbol{\Theta}^{(A)},\boldsymbol{\Phi}^{(A)}\right) \propto \underset{a\in\{1,\ldots,A\}}{\arg\max} \prod_{i=1}^{\tilde{N}} \sum_{t=1}^{T^{(A)}} \theta_{at}^{(A)} \phi_{t\tilde{w}_i}^{(A)} \tag{15}$$

This method does not require any topic inference in the classification phase, because the author topic distributions $\boldsymbol{\Theta}^{(A)}$ and topic word distributions $\boldsymbol{\Phi}^{(A)}$ are already inferred at training time. It is worth noting that we use the log of this probability for reasons of numerical stability.

As mentioned at the beginning of this section, we expect AT-P to outperform AT-SVM because AT-P relies directly on the probabilistic structure of the AT model. In addition, AT-P has the advantage of not requiring any topic inference in the classification phase.

We also performed preliminary experiments with a method that: (1) assumes that the test text was co-written by all the candidate authors, (2) infers the word-to-author assignments for the test text, and (3) returns the author that was attributed the most words. However, we found that this method performs poorly in comparison with other AT-based approaches in three-way authorship attribution. In addition, this method was too computationally expensive to run in cases with many authors, as it requires iterating through all the authors for every test text word in each sampling iteration.

4.4 Methods Based on AT-FA

AT-FA is the same model as AT, but it is run with the preprocessing step of adding an additional fictitious author to each training document. Hence, different constraints apply to AT-FA in the classification phase. This is because in this phase, we cannot conserve AT-FA's assumption that all the texts are written by a real author together with a fictitious author, since we do not know who wrote the test text. Hence, if we were to assume that the real author is a previously unknown author, as done for AT, we would have no way of telling the previously unknown author from the fictitious author, because they are both unique to the test text. We consider two possible ways of addressing this:

1. Assume that the test text was written only by a real, previously unknown, author (without a fictitious author), and infer this author's topic distribution $\tilde{\boldsymbol{\theta}}^{(A)}$ (as in AT).
2. For each training author a, assume that the test text was written by a together with a fictitious author f_a and infer the fictitious author's topic distribution $\tilde{\boldsymbol{\theta}}_{f_a}^{(A)}$. This results in a set of fictitious author topic distributions, each matching a training author.

Although the second alternative may appear more attractive because it does not violate the fictitious author assumption of AT-FA, we cannot use it with the dimensionality reduction method (AT-FA-SVM, as described in the following section), as this method requires inferring the topic distribution of the previously unknown author $\tilde{\boldsymbol{\theta}}^{(A)}$.

4.4.1 Dimensionality Reduction: AT-FA-SVM. AT-FA yields a topic distribution for each training document (i.e., the topic distribution of the fictitious author associated with the document), and a topic distribution for each real author (all the distributions are over the same topic set). We convert each training document to the concatenation of these two distributions, and use this concatenation as input to the SVM component. In the classification phase, we assume that the test text was written by a single previously unknown author, and represent the test text as the concatenation of the inferred topic distribution $\tilde{\boldsymbol{\theta}}^{(A)}$ to itself.

It is worth noting that our DADT model offers a more elegant solution than concatenating the same distribution to itself, because DADT differentiates between author topics and document topics—a distinction that AT-FA attempts to capture through fictitious authors. Hence, we expect the DADT-SVM approach, which we define in Section 4.5, to perform better than AT-FA-SVM. Nonetheless, we also experiment with AT-FA-SVM for the sake of completeness.

4.4.2 Probabilistic: AT-FA-P. For the probabilistic approach, we consider two variants, matching the two alternatives outlined earlier.

1. *AT-FA-P1.* This variant is identical in the classification phase to AT-P—it returns the author that maximizes the probability of the test text's words according to Equation (15), assuming that the test text was *not* co-written by a fictitious author.
2. *AT-FA-P2.* This variant performs the following steps for each author a: (1) assume that the test text was written by a and a fictitious author f_a; (2) infer the topic distribution of the fictitious author $\tilde{\boldsymbol{\theta}}^{(A)}_{f_a}$; (3) calculate the probability of the test text words under the assumption that it was written by a and f_a, and given the inferred $\tilde{\boldsymbol{\theta}}^{(A)}_{f_a}$; and (4) return the author for whom the probability of the test text words is maximized:

$$\underset{a\in\{1,\ldots,A\}}{\arg\max}\; p\left(\tilde{\boldsymbol{w}}|\tilde{\boldsymbol{a}},\tilde{\boldsymbol{\theta}}^{(A)}_{f_a},\boldsymbol{\Theta}^{(A)},\boldsymbol{\Phi}^{(A)}\right) \propto \underset{a\in\{1,\ldots,A\}}{\arg\max} \prod_{i=1}^{\tilde{N}} \sum_{t=1}^{T^{(A)}} \left(\theta^{(A)}_{at}\phi^{(A)}_{t\tilde{w}_i} + \tilde{\theta}^{(A)}_{f_a t}\phi^{(A)}_{t\tilde{w}_i}\right) \quad (16)$$

where $\tilde{\boldsymbol{a}} = \{a, f_a\}$ is the test text's set of authors.
The problem with this approach is that it is too computationally expensive to use on data sets with many candidate authors, as it requires running a separate inference procedure for each author. Nonetheless, in cases where AT-FA-P2 can be run, we expect it to perform better than AT-FA-P1 because it does not violate the fictitious author assumption of AT-FA.

4.5 Methods Based on DADT

4.5.1 Dimensionality Reduction: DADT-SVM. DADT yields a document topic distribution $\boldsymbol{\theta}^{(D)}_d$ for each document d, and an author topic distribution $\boldsymbol{\theta}^{(A)}_a$ for each author a. Similarly to AT-FA-SVM, we convert each training document to the concatenation of these two distributions, and use this concatenation as input to the SVM component.

In contrast to AT-FA, DADT's document topic distributions are defined over a topic set that is disjoint from the author topic set. This makes it possible to assume that the test text was written by a previously unknown author, and obtain the test text's document distribution $\tilde{\boldsymbol{\theta}}^{(D)}$ together with the previously unknown author's topic distribution $\tilde{\boldsymbol{\theta}}^{(A)}$ (following the procedure described in Section 3.4.2). As in the training phase, test texts are represented as the concatenation of these two distributions.

We expect DADT-SVM to outperform AT-FA-SVM, because we are able to maintain the assumptions of DADT in the classification phase, which we cannot do in AT-FA-SVM. Further, DADT-SVM should perform better than AT-SVM, because DADT-SVM accounts for differences between individual documents, whereas AT-SVM represents each author using a single training instance. Hypothesizing about the expected performance of DADT-SVM in comparison to LDA-SVM is harder: We expect performance to be corpus-dependent to a certain degree—in data sets where differences between individual documents are important, LDA-SVM may have an advantage, as all the words are allocated to document topics. On the other hand, in data sets where the differences between authors are more important, DADT-SVM may outperform LDA-SVM because it represents the authors explicitly.

4.5.2 Probabilistic: DADT-P. This method assumes that the test text was written by a previously unknown author, infers the test text's document topic distribution $\tilde{\boldsymbol{\theta}}^{(D)}$ and the author/document topic ratio $\tilde{\pi}$, and returns the most probable author according to the following equation:

$$\underset{a\in\{1,\ldots,A\}}{\arg\max}\; p\left(\tilde{a}=a|\tilde{\boldsymbol{w}},\tilde{\pi},\tilde{\boldsymbol{\theta}}^{(D)},\boldsymbol{\theta}_a^{(A)},\boldsymbol{\Phi}^{(D)},\boldsymbol{\Phi}^{(A)},\chi_a\right)\propto \tag{17}$$

$$\underset{a\in\{1,\ldots,A\}}{\arg\max}\;\chi_a\prod_{i=1}^{\tilde{N}}\left(\tilde{\pi}\sum_{t=1}^{T^{(A)}}\theta_{at}^{(A)}\phi_{t\tilde{w}_i}^{(A)}+(1-\tilde{\pi})\sum_{t=1}^{T^{(D)}}\tilde{\theta}_t^{(D)}\phi_{t\tilde{w}_i}^{(D)}\right)$$

It is worth noting that in preliminary experiments, we found that an alternative approach that avoids sampling $\tilde{\pi}$ and $\tilde{\boldsymbol{\theta}}^{(D)}$ by setting $\tilde{\pi}=1$ yields poor performance, probably because it "forces" all the words to be author words, including words that are very likely to be document words. In addition, we found that following an approach where $\tilde{\pi}$ and $\tilde{\boldsymbol{\theta}}^{(D)}$ are sampled separately for each author (similarly to AT-FA-P2) yields comparable performance to sampling only once by following the previously-unknown author assumption. However, the former approach is too computationally expensive to run on data sets with many candidate authors. Hence, we present only the results obtained with the approach that performs sampling only once.

5. Evaluation

This section presents the results of our evaluation. We first describe the data sets we used (Section 5.1) and our experimental setup (Section 5.2), followed by the results of our experiments on the Judgment and PAN'11 data sets (Section 5.3). Then, we present the results of a more restricted set of experiments on the larger IMDb62, IMDb1M, and Blog data sets (Section 5.4) and summarize our key findings (Section 5.5).

5.1 Data Sets

We experimented with five data sets: Judgment, PAN'11, IMDb62, IMDb1M, and Blog. Judgment, IMDb62, and IMDb1M were collected and introduced by us, and are freely available for research use (Judgment can be downloaded from `www.csse.monash.edu.au/research/umnl/data`, and IMDb62 and IMDb1M are available upon request). The two other data sets were introduced by other researchers, are publicly available, and were used to facilitate comparison between our methods and previous work. Table 3 presents some data set statistics.

5.1.1 Judgment. The Judgment data set contains judgments by three judges who served on the Australian High Court from 1913 to 1975: Dixon, McTiernan, and Rich (abbreviated to D, M, and R, respectively, in Table 3). We created this data set to verify rumors that Dixon ghost-wrote some of the judgments attributed to McTiernan and Rich (Seroussi, Smyth, and Zukerman 2011). This data set is an example of a traditional authorship attribution data set, as it contains only three authors who wrote relatively long texts in a formal language. In this article, we only use judgments with undisputed authorship, which were written in periods when only one of the three judges served on the High Court (Dixon's 1929–1964 judgments, McTiernan's 1965–1975 judgments, and Rich's 1913–1928 judgments). We removed numbers from the texts to ensure that dates cannot be used to discriminate between judges. We also removed quotes to ensure that the classifiers take into account only the actual authors' language use (removal was done automatically by matching regular expressions for numbers and text in quotation marks). Because all three judges dealt with various topics, it is likely that successful methods would have to consider each author's style, rather than rely solely on content features in the texts.

As Table 3 shows, the Judgment data set contains the smallest number of authors of the data sets we considered, but these authors wrote more texts than the average author in PAN'11, IMDb1M, and Blog. Judgments are also substantially longer than the texts in all the other data sets, which should make authorship attribution on the Judgment data set relatively easy.

Table 3
Data set statistics.

	Judgment	PAN'11	IMDb62	IMDb1M	Blog
Authors	3	72	62	22,116	19,320
Texts	1,342	Trn: 9,335 Vld: 1,296 Tst: 1,300	79,550	271,625	678,161
Texts per author mean (stddev)	D: 902 M: 253 R: 187	Trn: 129.7 (139.3) Vld: 19.9 (19.0) Tst: 20.3 (18.9)	1283.1 (685.8)	12.3 (92.1)	35.1 (105.0)
Tokens per text mean (stddev)	D: 2,858.6 (2,456.9) M: 1,310.7 (1,248.4) R: 783.0 (878.5)	Trn: 60.8 (109.4) Vld: 65.3 (98.9) Tst: 71.0 (115.1)	281.8 (234.8)	124.2 (166.6)	248.4 (510.8)

5.1.2 PAN'11. The PAN'11 data sets were introduced as part of the PAN 2011 competition (available from pan.webis.de) (Argamon and Juola 2011). These data sets were extracted from the Enron e-mail corpus (www.cs.cmu.edu/~enron), and were designed to emulate closed-class and open-class authorship attribution and authorship verification scenarios (Section 2). These data sets represent authorship attribution scenarios that may arise in computer forensics, such as the case noted by Chaski (2005), where an employee who was terminated for sending a racist e-mail claimed that any person with access to his computer could have sent the e-mail.

In our experiments, we used the largest PAN'11 data set, with e-mails by 72 authors. Unlike the other data sets we used, this data set is split into training, validation, and testing subsets (abbreviated to Trn, Vld, and Tst, respectively, in Table 3). We focused on the closed-class problem, using the validation and testing sets that contain texts only by training authors. The only change we made to the original data set was dropping two training and two validation texts that were automatically generated, which were detected by length and content. This had a negligible effect on method accuracy, but made the statistics in Table 3 more representative of the data (e.g., the mean count of tokens per text is 65.3 in the validation set without the two automatically generated texts, compared with 338.3 in the full validation set).

Using this data set allows us to test our methods on short and informal texts with more authors than in traditional authorship attribution. As Table 3 shows, the PAN'11 data set contains the shortest texts of the data sets we considered. This fact, together with the training/validation/testing structure of the data set, make it possible to run many experiments on this data set before moving on to larger data sets.

5.1.3 IMDb62. IMDb62 contains 62,000 movie reviews and 17,550 message board posts by 62 prolific users of the Internet Movie database (IMDb, www.imdb.com). We introduced this data set (Seroussi, Zukerman, and Bohnert 2010) to test our author-aware polarity inference approach (Section 6.2). Each user wrote 1,000 reviews (sampled from their full set of reviews), and a variable number of message board posts, which are mostly movie-related, but may also be about television, music, and other topics. This data set 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. Unlike the other data sets of informal texts, IMDb62 consists only of prolific authors, allowing us to test our approach in a scenario where training data is plentiful.

5.1.4 IMDb1M. Although the IMDb62 data set is useful for testing our methods on small- to medium-scale problems, it cannot be seen as an adequate representation of large-scale problems. This is especially relevant to the task of rating prediction, in which typical data sets contain thousands of users (Section 6.3). Hence, we created IMDb1M by randomly generating one million valid IMDb user IDs and downloading the reviews and message board posts written by these users (Seroussi, Bohnert, and Zukerman 2011). Unfortunately, most of the randomly generated IDs led to users who submitted neither reviews nor posts—we found that about 5% of the entire user population submitted posts, and less than 3% wrote reviews. After filtering out users who have not submitted any rated reviews, we were left with 22,116 users. These users, who make up the IMDb1M data set, submitted 204,809 posts and 66,816 rated reviews.

IMDb1M can be seen as complementary to the IMDb62 data set, as IMDb62 allows us to test scenarios in which the user population is made up of prolific users, whereas

IMDb1M contains a more varied sample of the population. However, because we did not impose a minimum threshold on the number of reviews or posts, the IMDb1M population is very challenging as it includes many users with few texts (e.g., about 56% of the users in IMDb1M wrote only one text). It is worth noting that three users appear in both IMDb62 and IMDb1M. In IMDb62 these three users authored 3,000 reviews and 268 posts in total (about 4.8% of the total number of reviews and 1.5% of the posts), and in IMDb1M they authored 5,695 reviews and 358 posts (about 8.5% of the reviews and 0.2% of the posts). The difference in the number of reviews is due to the sampling we performed when we created IMDb62, and the difference in the number of posts is due to the time difference between the creation of the two data sets.

5.1.5 Blog. The Blog data set is the largest data set we consider, containing 678,161 blog posts by 19,320 authors (available from `u.cs.biu.ac.il/~koppel`). It was created by Schler et al. (2006) to learn about the relation between language use and demographic characteristics, such as age and gender. We use this data set to test how our authorship attribution methods scale to handle thousands of authors. As blog posts can be about any topic, this data set is less restricted than the Judgment, PAN'11, and IMDb data sets. Further, the large number of authors ensures that every topic is likely to interest at least several authors, meaning that methods that rely only on content are unlikely to perform as well as methods that also take author style into account.

5.2 Experimental Setup

We used different experimental setups, depending on the data set. PAN'11 experiments followed the setup of the PAN'11 competition (Argamon and Juola 2011): We trained all the methods on the given training data set, tuned the parameters according to results obtained for the given validation data set, and ran the tuned methods on the given testing data set. For all the other data sets we utilized ten-fold cross validation. In all cases, we report the overall classification accuracy, that is, the percentage of test texts correctly attributed to their author. Statistically significant differences are reported when $p < 0.05$ according to McNemar's test (when reporting results in a table, the best result for each column is in boldface, and several boldface results mean that the differences between them are not statistically significant).

In our experiments, we used the L2-regularized linear SVM implementation of LIBLINEAR (Fan et al. 2008), which is well suited for large-scale text classification. We experimented with cost parameter values from the set $\{\ldots, 10^{-1}, 10^{0}, 10^{1}, \ldots\}$, until no accuracy improvement was obtained (starting from $10^0 = 1$ and going in both directions). We report the results obtained with the value that yielded the highest accuracy, which gives an optimistic estimate for the performance of the Token SVM baseline.

We used collapsed Gibbs sampling to train all the topic models, running four chains with a burn-in of 1,000 iterations. In the Judgment, PAN'11, and IMDb62 experiments, we retained eight samples per chain with a spacing of 100 iterations. In the IMDb1M and Blog experiments, we retained one sample per chain due to runtime constraints. Because we cannot average topic distribution estimates obtained from training samples due to topic exchangeability (Steyvers and Griffiths 2007), we averaged the probabilities calculated from the retained samples. In the dimensionality reduction experiments, we used the topic distributions from a single training sample to ensure that the number of features is substantially reduced (an alternative approach would be to use the concatenation of all the samples, but this may result in a large number of features, and employing this alternative approach did not improve results in preliminary experiments).

For test text sampling, we used a burn-in of 10 iterations and averaged the parameter estimates over the next 10 iterations in a similar manner to the procedure used by Rosen-Zvi et al. (2010). We found that these settings yield stable results across different random seed values.

To enable a fair comparison between the topic-based methods and the Token SVM baseline, all methods were trained on the same token representations of the texts. In most experiments, we did not apply any filters and simply used all the tokens as they appear in the text. In some cases, as indicated throughout this section, we either retained only stopwords or discarded the stopwords in a preprocessing step that was applied before running the methods. This allowed us to obtain rough estimates of the effect of considering only style words, considering only content words, and considering both style and content. However, we note that this is only a crude way of separating style and content, because some stopwords may contain content clues, whereas some words that do not appear in the stopword list may be seen as indicators of personal style, regardless of content.

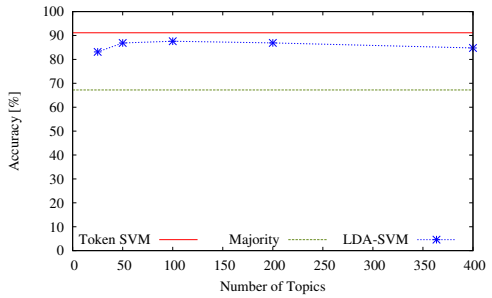
We found that the number of topics has a large impact on performance, and the effect of other configurable parameters is smaller (Section 3.1). Hence, we used symmetric topic priors, setting all the elements of $\boldsymbol{\alpha}^{(D)}$ and $\boldsymbol{\alpha}^{(A)}$ to $\min\{0.1, 5/T^{(D)}\}$ and $\min\{0.1, 5/T^{(A)}\}$, respectively. For all models, we set $\beta_w = 0.01$ for each word w as the base measure for the prior of words in topics. Because 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. Unless otherwise specified, we set $\epsilon = 0.009$, based on tuning experiments on Judgment and PAN'11 (Section 5.3). Similarly, we set $\delta^{(D)} = 1.222$ and $\delta^{(A)} = 4.889$ for DADT, based on the same experiments. In addition, we set $\eta_a = 1$ for each author a, yielding smoothed estimates for the corpus distribution of authors $\boldsymbol{\chi}$.

5.3 Experiments on Small Data Sets

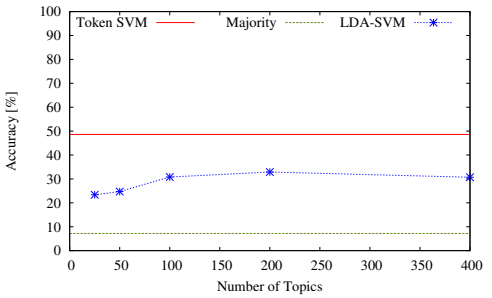
In this section, we present the results of our experiments on the Judgment data set, which contains judgments by three judges, and on the PAN'11 data set, which contains e-mails by 72 authors. Authorship attribution on the PAN'11 data set is more challenging than on the Judgment data set, because PAN'11 texts are shorter than judgments, and some of the PAN'11 authors wrote only a few e-mails. We first present the results obtained with LDA, followed by the results obtained with AT (with and without fictitious authors), and with our DADT-based methods, which yielded the best performance. We end this section with experiments that explore the effect of applying stopword filters to the corpus in a preprocessing step. These experiments demonstrate that our DADT-based approach models authorship indicators other than content words.

As discussed in Section 5.2, we ran ten-fold cross validation on the Judgment data set. On PAN'11, we tuned the methods on the validation subset and report the results obtained on the testing subset with the settings that yielded the best validation results (i.e., each method was run multiple times on the validation subset and only once on the testing subset). We present some tuning results together with testing results to illustrate the behavior of the various methods. It is worth noting that for most methods, PAN'11 testing results are better than the best validation results. This may be because on average testing texts are about 10% longer than validation texts (Section 5.1.2).

5.3.1 LDA. Figure 5 presents the results of the LDA experiment, with Judgment results in Figure 5a, and PAN'11 validation and testing results in Figures 5b and 5c, respectively.



(a) Judgment data set



(b) PAN'11 validation

Method	$T^{(D)}$	**Accuracy**
Majority	—	7.15%
Token SVM	—	**53.31%**
LDA-SVM	200	31.92%

(c) PAN'11 testing

Figure 5
LDA results (data sets: Judgment and PAN'11).

On the Judgment data set, the best performance obtained by training an SVM classifier on LDA topic distributions (LDA-SVM with 100 topics) was somewhat worse than that obtained by training directly on tokens (Token SVM), but was still much better than a majority baseline (the differences between LDA-SVM and both the Token SVM and majority baselines are statistically significant in all cases). This indicates that although some authorship indicators are lost when using LDA for dimensionality reduction, many are retained despite the fact that LDA's document representations are much more compact than the raw token representations.

The ranking of methods on PAN'11 is similar to the ranking on the Judgment data set, though on Judgment the difference between LDA-SVM and Token SVM is much smaller. The reason for this difference may be that LDA does not consider authors in the model-building stage. Although this had a relatively small effect on performance in the three-way judgment attribution scenarios, it appears that accounting for authors is important in scenarios with many authors. As the rest of this article deals with such scenarios, we decided not to use LDA for modeling authors in subsequent sections.

5.3.2 AT. Figure 6 presents the results of the AT experiment, with Judgment results in Figure 6a and PAN'11 validation and testing results in Figures 6b and 6c, respectively.

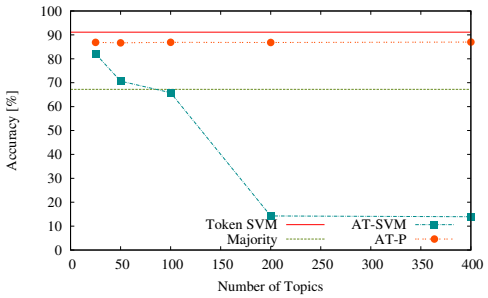
In contrast to LDA-SVM, AT-SVM was very sensitive to the number of topics. This is probably because AT-SVM's dimensionality reduction is more radical than LDA-SVM's: In AT-SVM, each document is reduced to the same distribution over author topics because AT does not model individual documents (Section 4.3). Notably, AT-SVM's performance was very poor when 200 and 400 topics were used, possibly because the more fine-grained topic distributions yielded by using more topics resulted in sparser author representations (where some topics were allocated only a few words), which may have caused the SVM component to overfit. This trend is more pronounced in the Judgment results than in the PAN'11 results: On Judgment, AT-SVM with 200 and

400 topics yielded poorer results than the majority baseline, probably because the effect of sparsity is larger when considering three authors than when modeling 72 authors.

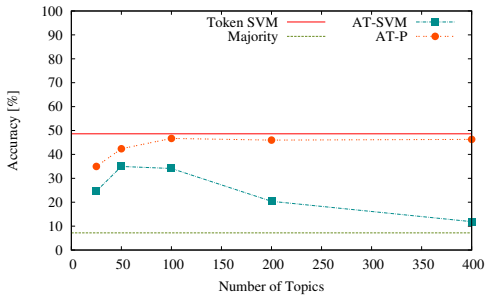
The probabilistic AT-P method significantly outperformed AT-SVM and the majority baseline on both data sets. Although AT-P performed comparably to Token SVM on the PAN'11 data set, it was significantly outperformed by Token SVM on the Judgment data set. Nonetheless, these results indicate that AT captures many of the indicators required for authorship attribution. This is despite the fact that AT was not designed with authorship attribution in mind. Hence, it represents each author with a single distribution over topics while ignoring differences and similarities between documents (which may be important for the authorship attribution task). This stands in contrast to the Token SVM baseline, which attempts to build a document-based model that is optimized for the classification goal of authorship attribution (Section 4.1).

5.3.3 AT-FA. Figure 7 presents the results of the AT-FA experiment, with Judgment results in Figure 7a and PAN'11 validation and testing results in Figures 7b and 7c, respectively.

On both data sets, the highest accuracy yielded by AT-FA-SVM and AT-FA-P1 was significantly lower than that obtained by the corresponding methods in the AT case *without* fictitious authors (AT-SVM and AT-P, respectively). This may seem surprising, because the only difference between AT and AT-FA is the addition of a fictitious author for each document, which was shown to improve AT's ability to predict unseen portions of documents (Rosen-Zvi et al. 2010). However, the reason for AT-FA-SVM and AT-FA-P1's poor performance may be that they do not conserve the underlying assumption of fictitious authors in the classification stage, i.e., they do not assume that the test text was written by a fictitious author together with a previously unseen author (Section 4.4). This is probably also the reason why the probabilistic AT-FA-P2 signifi-



(a) Judgment data set

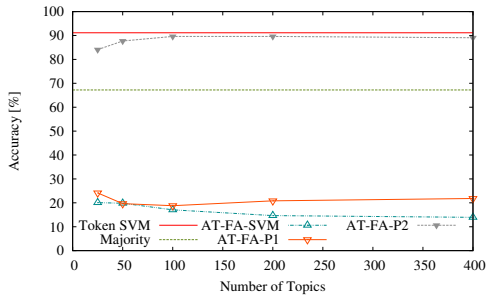


(b) PAN'11 validation

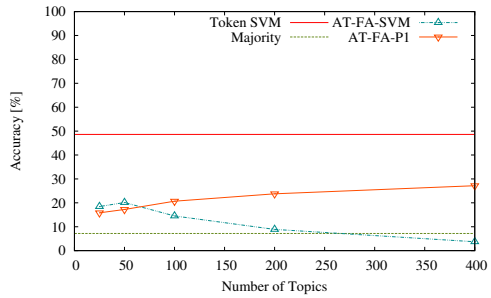
Method	$T^{(A)}$	**Accuracy**
Majority	—	7.15%
Token SVM	—	**53.31%**
AT-SVM	50	39.23%
AT-P	100	**53.08%**

(c) PAN'11 testing

Figure 6
AT results (data sets: Judgment and PAN'11).



(a) Judgment data set



(b) PAN'11 validation

Method	$T^{(A)}$	**Accuracy**
Majority	—	7.15%
Token SVM	—	**53.31%**
AT-FA-SVM	50	23.15%
AT-FA-P1	400	29.69%

PAN'11 testing

Figure 7
AT-FA results (data set: Judgment and PAN'11).

cantly outperformed AT-FA-P1 by a large margin on the Judgment data set—AT-FA-P2 conserves the fictitious author assumption, whereas AT-FA-P1 ignores it (we did not run the AT-FA-P2 method on PAN'11 because it requires running a separate sampling chain for each candidate author and test text, which makes it too computationally expensive to run in cases with many candidate authors and test texts).

When comparing AT-FA-P2 to the baselines (on Judgment), we see that it was significantly outperformed by Token SVM for all topic numbers, but yielded significantly better performance than the majority baseline. Despite the fact that AT-FA-P2 was outperformed by Token SVM, the margin was not large when enough topics were used (AT-FA-P2 yielded its best accuracy of 89.60% with 100 topics, in comparison with Token SVM's accuracy of 91.15%). This indicates that representing both documents and authors in the topic model may have advantages in terms of authorship attribution. This further motivates the use of our DADT model, which considers documents and authors without requiring the preprocessing step of adding fictitious authors.

5.3.4 DADT. Figure 8a presents the results of the DADT experiment on the Judgment data set, obtained with 10 author topics, 90 document topics, and prior settings of $\delta^{(D)} = 1.222$, $\delta^{(A)} = 4.889$, and $\epsilon = 0.009$ (other parameter settings are discussed subsequently). These results are compared to the baselines (majority and Token SVM), and to the best topic-based result obtained on this data set thus far (by AT-FA-P2 with 100 topics). As we can see, the best DADT-based result was obtained with the probabilistic DADT-P method, which significantly outperformed all the other methods. This demonstrates the effectiveness of our DADT model in capturing author characteristics that are relevant to authorship attribution.

Notably, DADT-SVM yielded significantly poorer results than DADT-P. DADT-SVM's relatively weak performance may be because its use of document topics

Method	Accuracy
Majority	67.21%
Token SVM	91.15%
AT-FA-P2	89.60%
DADT-SVM	85.49%
DADT-P	**93.64%**

(a) Tuned DADT methods

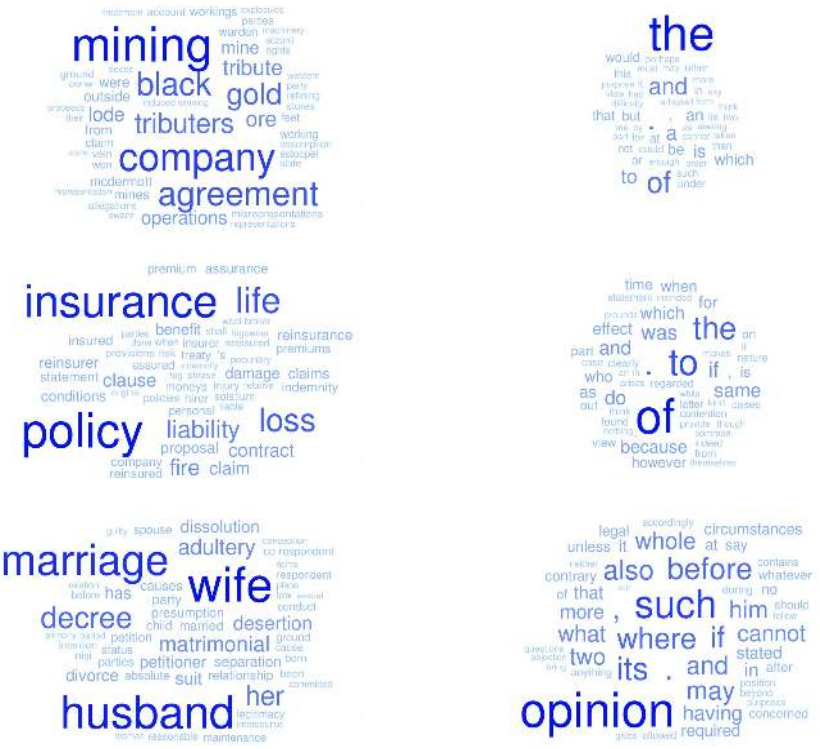
$T^{(D)}$	$T^{(A)}$	$\delta^{(D)}$	$\delta^{(A)}$	ϵ	Accuracy
90	10	1	1	0	**93.81%**
90	10	1.222	4.889	0	**93.49%**
90	10	1.222	4.889	0.009	**93.64%**
50	50	1.222	4.889	0.009	92.88%
10	90	1.222	4.889	0.009	88.62%

(b) DADT-P tuning

Figure 8
DADT results (data set: Judgment).

introduces noise that causes the SVM component to underperform, as DADT's document topics are not expected to be indicative of authorship.

The separation of document words from author words that is obtained by using DADT on the Judgment data set is illustrated by Figure 9, which presents three document topics and three author topics in word-cloud form. The top 50 tokens from each topic are shown, where the size and shade of each token indicates its frequency in the topic. This anecdotal sample of topics reflects the general trend that we noticed in this data set, where document topics represent different types of cases, and the top tokens in author topics do not carry content information and are dominated by stopwords



(a) Document topics (b) Author topics

Figure 9
DADT topic examples.

and punctuation (LDA and AT topics were similar to DADT's author topics due to the prevalence of stopwords and the lack of document–author separation in these models). This trend is in line with what we expected, because all three judges handled cases of different types, and thus content words are unlikely to carry enough information to adequately represent the judges. As discussed in Section 3.1.3, this separation of content and style is corpus-dependent and is expected to occur only in cases where content is independent of author identity. Indeed, we did not observe such a clear separation in our experiments on other data sets.

Our choice of DADT settings reflects the following insights:

- We used 100 topics overall based on the results of the other topic-based methods, which showed that good results are obtained with this number of overall topics. We chose the 90/10 document/author topic split because in the case of the Judgment data set, DADT models only three authors who wrote many documents.
- Setting $\delta^{(D)} = 1.222$ and $\delta^{(A)} = 4.889$ encodes our prior belief that the portion of each document that is composed of author words is 80% on average, with 15% standard deviation (obtained as described in Section 3.1.3).
- Setting $\epsilon = 0.009$ encodes our prior belief that stopword choice is more likely to be influenced by the identity of the author than by the content of the documents (Section 5.2).

Somewhat surprisingly, these settings did not have a large effect on the performance of the methods in most cases. This is demonstrated by the results presented in Figure 8b, which were obtained by varying the values of these parameters and running the DADT-P method. As Figure 8b shows, the results obtained with a setting of $\delta^{(D)} = \delta^{(A)} = 1$, which can be seen as encoding no strong prior belief about the document/author word balance in each document (it is equivalent to setting a uniform prior on this balance), were comparable to the results obtained with $\delta^{(D)} = 1.222$ and $\delta^{(A)} = 4.889$. Likewise, changing ϵ from 0 to 0.009 only had a minor effect on the results. The only setting that made a relatively large difference is the document/author topic split: Changing it from 90/10 to 10/90 yielded poorer results. However, the 50/50 split yielded close results to the 90/10 split, which shows that in this case, the document/author topic split setting is only sensitive to relatively large variations.

It is likely that performing an exhaustive grid search for the optimal parameter settings for each method would allow us to obtain somewhat improved results. However, such a search would be computationally expensive, as the model needs to be retrained and tested for each fold, parameter set, and method. Therefore, we decided to present the results obtained with the non-optimized settings, which are sufficient to demonstrate the merits of our DADT approach, as DADT-P outperformed all the other methods discussed so far.

On PAN'11, we ran the DADT experiments with 100 topics overall, as this number of topics yielded the best topic-based results of the models and methods whose results we presented thus far (AT-P with 100 topics yielded the best results of the methods based on LDA, AT, and AT-FA). Figure 10b shows the results of tuning DADT's settings and running DADT-P on the PAN'11 validation set. The PAN'11 tuning experiment shows a clearer picture in terms of accuracy differences between different parameter settings than the Judgment experiments. Specifically, when we used uninformed uniform priors on the document/author word split ($\delta^{(D)} = \delta^{(A)} = 1$), and the same word-in-topic priors for both document and author words ($\epsilon = 0$), the obtained accuracy was comparable to AT-P's accuracy. On the other hand, setting $\delta^{(D)} = 1.222$ and $\delta^{(A)} = 4.889$, which encodes our prior belief that on average 80% (with a standard deviation of 15%)

of each document is composed of author words, significantly improved performance. Setting $\epsilon = 0.009$ to encode our prior knowledge that stopwords are indicators of authorship yielded an additional improvement. Finally, the last two results in Figure 10b demonstrate the importance of having enough topics to model the authors: Accuracy dropped by about 4 percentage points when we used 50 author topics and 50 document topics, and by about 24 percentage points when we used only 10 author topics and 90 document topics, rather than 90 author topics and 10 document topics. This leads us to conjecture that it would be beneficial to pursue a future extension that learns 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.

Figure 10a presents the PAN'11 results obtained with the DADT-based methods, using the best setting from Figure 10b: 10 document topics, 90 author topics, $\delta^{(D)} = 1.222$, $\delta^{(A)} = 4.889$, and $\epsilon = 0.009$. As Figure 10a shows, DADT-P, which obtained the best performance of all the methods tested in this section, is the only method that outperformed Token SVM. This implies that our DADT model is the most suitable of the models we considered for capturing patterns in the data that are important for authorship attribution, at least in scenarios that are similar to the PAN'11 case.

DADT-P's testing result is comparable to the third-best accuracy (out of 17) obtained in the PAN'11 competition (Argamon and Juola 2011) (competitors were ranked according to macro-averaged and micro-averaged precision, recall, and F1; the micro-averaged measures are all equivalent to the accuracy measure in this case, because each of the test texts is assigned to a single candidate author). However, to the best of our knowledge, DADT-P 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, whereas 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). In addition, 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 directions for future work.

5.3.5 Testing the Effect of Stopwords. The results reported up to this point were all obtained by running the methods on document representations that include all the tokens. As discussed in Section 5.2, discarding or retaining stopwords provides a crude way of separating style from content. We ran a set of experiments where we either discarded

Method	**Accuracy**
Majority	7.15%
Token SVM	53.31%
AT-P	53.08%
DADT-SVM	39.69%
DADT-P	**59.38%**

(a) Tuned DADT methods (testing subset)

$T^{(D)}$	$T^{(A)}$	$\delta^{(D)}$	$\delta^{(A)}$	ϵ	**Accuracy**
10	90	1	1	0	48.53%
10	90	1.222	4.889	0	53.40%
10	90	1.222	4.889	0.009	**54.86%**
50	50	1.222	4.889	0.009	50.31%
90	10	1.222	4.889	0.009	30.48%

(b) DADT-P tuning (validation subset)

Figure 10
DADT results (data set: PAN'11).

Table 4
Stopword experiment results (data sets: Judgment and PAN'11).

Method	Judgment			PAN'11 Testing		
	All words	Discard stopwords	Retain only stopwords	All words	Discard stopwords	Retain only stopwords
Majority	67.21%	67.21%	67.21%	7.15%	7.15%	7.15%
Token SVM	91.15%	86.18%	**92.76%**	53.31%	46.46%	**28.38%**
DADT-P	**93.64%**	**89.28%**	90.85%	**59.38%**	**54.69%**	18.54%

stopwords in a preprocessing step or retained only stopwords, and then ran the Token SVM baseline and the DADT-P method, which obtained the best performance when all the tokens were used (DADT was run with the same settings used to obtain the tuned results from the previous section).

The results of this experiment are presented in Table 4. As the results show, discarding stopwords caused the Token SVM baseline to yield poorer performance than when all the tokens were used, but retaining only stopwords significantly improved Token SVM's performance on Judgment and yielded a substantial drop in performance on PAN'11. Interestingly, this was not the case with DADT-P, where either discarding or retaining stopwords caused a statistically significant drop in performance in comparison with using all the tokens. The reason why DADT-P's performance dropped when only stopwords were used may be that DADT was designed under the assumption that all the tokens in the corpus are retained. However, we are encouraged by the fact that DADT-P's performance drop on Judgment was not very large when only stopwords were retained, as it indicates that DADT captures stylistic elements in the authors' texts.

Another encouraging result is that DADT-P yielded significantly better performance than Token SVM when using feature sets that included all the tokens or all the tokens without stopwords. DADT-P appears to harness the extra information from non-stopword tokens more effectively than Token SVM, despite the fact that such tokens tend to occur less frequently in the texts than stopwords. Further, the vocabulary size of these two feature sets is larger than that of the stopword-only feature set, which suggests that DADT-P is more resilient to noise than Token SVM.

It is worth noting that some content-independent information is lost when only stopwords are retained. For example, the phrase "in my opinion" appears in texts by all three authors in the Judgment data set, but is used more frequently by McTiernan (it occurs in about 82% of his judgments) than by Dixon (69%) or Rich (58%). As the frequency of this phrase is apparently dependent on author style and independent of the specific content of a given judgment, it is probably safe to assume that it would be beneficial to retain the word "opinion" (this is also evidenced by the dominance of this word in the third author topic in Figure 9). However, this word does not appear in our stopword list. This problem is more pronounced in the PAN'11 data set, where it appears that other words beyond stopwords are also indicative of authorship. For instance, Tanguy et al. (2011) used the openings and closings of the e-mails in the data set as separately weighted features. Openings can start with words such as "hello," "hi," "hey," and "dear," but only the first two words appear in our stopword list, meaning that even when only stopwords are retained some stylistic features are lost. These examples highlight the difficulties in extracting words that are truly content-independent—a problem that would be especially relevant when trying to adapt an authorship classifier

Table 5
Large-scale experiment results (data sets: IMDb62, IMDb1M, and Blog).

Method	IMDb62	IMDb1M	Blog (prolific)	Blog (full)
Majority	7.37%	3.00%	1.28%	0.62%
Token SVM	**92.52%**	43.85%	32.96%	24.13%
AT-P	89.62%	40.82%	37.59%	23.03%
DADT-P	91.79%	**44.23%**	**43.65%**	**28.62%**

trained on texts from one domain to texts from a completely different domain (this problem is beyond the scope of this study). A possible solution is to obtain corpus-specific stopwords—for example, by extracting a list of frequent words—but this gives rise to new problems, such as determining a frequency threshold. We decided not to pursue such a solution because the PAN'11 results show that improved performance is not guaranteed when only stopwords are retained, even when Token SVM is used. Hence, in the remainder of this article we use all the words, that is, we neither discard stopwords nor retain only stopwords.

5.4 Experiments on Large Data Sets

In this section, we report the results of our experiments on the IMDb62, IMDb1M, and Blog data sets. Both IMDb data sets contain movie reviews and message board posts, with IMDb62 consisting of texts by 62 prolific authors (with at least 1,000 texts each), and IMDb1M consisting of texts by 22,116 authors, who are mostly non-prolific. The Blog data set contains blog posts by 19,320 authors, and is the largest of the data sets we considered in terms of token count—it contains about 168 million tokens, whereas IMDb62 and IMDb1M contain about 22 and 34 million tokens, respectively. In addition to running experiments on the full Blog data set, we considered a subset that contains all the texts by the 1,000 most prolific authors (this subset contains about 69 million tokens overall in 332,797 posts—about 49% of the posts in the full Blog data set).

Due to resource constraints, we performed a more restricted set of experiments on IMDb62, IMDb1M, and Blog than on the Judgment and PAN'11 data sets (which contain about 3 and 0.74 million tokens, respectively). We ran only the Token SVM baseline, AT-P, and DADT-P, as these methods yielded the best performance in the PAN'11 experiments. We set the overall number of topics of AT and DADT to 200 topics for IMDb62, and 400 topics for IMDb1M and Blog. We set DADT's document/author topic split to 50/150 for IMDb62 and 50/350 for IMDb1M and Blog, and used the prior setting that yielded the best PAN'11 results ($\delta^{(D)} = 1.222, \delta^{(A)} = 4.889$, and $\epsilon = 0.009$). As in the PAN'11 experiments, we determined the overall number of topics based on AT-P's performance with 25, 50, 100, 200, and 400 topics. The document/author topic splits we tested were 10/190, 50/150, and 100/100 for IMDb62, and 10/390, 50/350, and 100/300 for IMDb1M and Blog.

Table 5 shows the results of this set of experiments. As in our previous experiments, DADT-P consistently outperformed AT-P, which indicates that using disjoint sets of document and author topics yields author representations that are more suitable for authorship attribution than using only author topics. In contrast to the previous experiments, Token SVM outperformed DADT-P in one case: the IMDb62 data set. This may be because discriminative methods (such as Token SVM) tend to outperform generative

methods (such as DADT-P) in scenarios where training data is abundant (Ng and Jordan 2001), which is the case with IMDb62—it contains at least 900 texts per author in each training fold.

A notable result is that although all the methods yielded relatively low accuracies on the full Blog data set, the topic-based methods experienced a larger drop in accuracy than Token SVM when transitioning from the prolific author subset to the full data set. This may be because topic-based methods use a single model, making them more sensitive to the number of authors than Token SVM's one-versus-all setup that uses one model per author (this sensitivity may also explain why DADT-P outperformed Token SVM by a relatively small margin on IMDb1M). This result suggests a direction for future work in the form of an ensemble of Token SVM and DADT-P. The potential of this direction is demonstrated by the fact that a perfect oracle, which chooses the correct answer between Token SVM and DADT-P when they disagree, yields an accuracy of 37.36% on the full Blog data set.

5.5 Summary of Key Findings

In summary, we found that the DADT-based probabilistic approach (DADT-P) yielded strong performance on the five data sets we considered, outperforming the Token SVM baseline in four out of the five cases. We showed that DADT-P is more suitable for authorship attribution than methods based on LDA and AT (with or without fictitious authors), and than using DADT for dimensionality reduction. Although our results demonstrate that separating document words from author words is a good approach to authorship attribution, relying only on unigrams is a limitation (which is shared by LDA, AT, and DADT). We discuss ways of addressing this limitation in Section 7.

DADT's improved performance in comparison with methods based on LDA and AT comes at a price of more parameters to tune. However, the most important parameter is the number of topics—we found that the prior values that yielded good results on the small data sets also obtained good performance on the large data sets without further tuning. We offered a simple recipe to determine the number of topics for DADT-P: First run AT-P to find the overall number of topics (which is equivalent to running DADT-P without document topics), and then tune the document/author topic balance. As mentioned in Section 3.1.2, this procedure can be obviated by automatically learning the topic balance, which is left for future work.

6. Applications

This section presents three applications of topic-based author representations: identifying anonymous reviewers (Section 6.1), author-aware polarity inference (Section 6.2), and text-aware rating prediction (Section 6.3).

6.1 Reviewer Identification

AT and DADT can potentially be used to identify anonymous reviewers based on publicly available data—the reviewer list (which is commonly available), and the reviewers' published papers. The main question in this case is whether authorship markers learned from (often multi-authored) texts in one domain (the papers) can be used to classify single-authored texts from a related domain (the reviews).

To start answering this question, we considered a small conference track, which attracted 18 submissions that were each reviewed by two reviewers. We collected the bodies of 10 papers (without references, author names, acknowledgments, etc.) by each of the 18 reviewers that were listed in the proceedings, which resulted in a training corpus of 171 documents with 196 authors overall (some of the reviewers have co-authored papers with other reviewers). We omitted authors with only one paper, because their presence is equivalent to having fictitious authors, which may hurt performance (Section 5.3). This resulted in a total of 77 authors. Our test data set consisted of 19 reviews by the 9 reviewers who gave us permission to use their reviews.

We trained AT and DADT on the paper corpus under the setup described in Section 5.2, and used AT-P and DADT-P to classify the reviews. The best accuracy, 8/19, was obtained by DADT-P with 10 document topics and 90 author topics. The accuracy of AT-P (with 100 topics) was slightly worse, at 7/19. In addition, the correct reviewer appeared in the top-five list of probable authors for 15/19 of the reviews with DADT-P and 11/19 with AT-P (the list of probable authors included all 18 reviewers—we considered all the reviewers as candidates because this did not require using any private information and it made our experimental setup more realistic). We obtained better results by eliminating non-reviewers from the training corpus (thus training on the 171 documents with 18 authors overall). DADT-P required only 25 document topics and 25 author topics in this case, and its accuracy rose to 10/19 (AT-P again performed worse with an accuracy of 7/19). In 16/19 of the cases the correct reviewer appeared in DADT-P's top-five list, compared to 12/19 with AT-P.

These results were obtained on a very small data set. Still, they indicate that reviewer identification is feasible (note that it is unlikely that DADT-P's performance is only due to content words, as interest areas are often shared between reviewers). To verify this, a fully fledged study should be done on a corpus of reviews from a large conference, with a training corpus that includes each author's full body of publications (perhaps dropping very old publications, which we did not do). As far as we know, such a study is yet to be performed. The closest work we know of is by Nanavati et al. (2011), who considered the question of whether "insiders," who served as program committee members and thus had access to non-anonymous reviews, can use these reviews as training data to identify reviewers. Although they found that they could identify reviewers with high accuracy, the main limitation of their approach is that it relies on private data.

Nonetheless, we believe that reviewer anonymity needs to be addressed. One approach is to use tools that obfuscate author identity, as developed by, for example, Kacmarcik and Gamon (2006) and Brennan and Greenstadt (2009). However, as this may lead to an "arms race" between such tools and authorship analysis methods, perhaps the best approach is to forgo anonymity completely, as advocated by some researchers and editors (Groves 2010). This is an open question with no simple answers, but we hope that our results will help motivate the search for solutions.

6.2 Author-Aware Polarity Inference

Sentiment analysis deals with inferring people's sentiments and opinions from texts (Pang and Lee 2008; Liu and Zhang 2012). One of the main tasks in this field is **polarity inference**, where the goal is to infer the degree of positive or negative sentiment of texts (Pang and Lee 2008). Even though the way polarity is expressed often depends on the author, most of the work in this field ignores authors. We addressed this gap (Seroussi, Zukerman, and Bohnert 2010; Seroussi 2012) by introducing a framework

that considers authors when performing polarity inference, by combining the outputs of author-specific inference models in a manner that makes it possible to consider author similarity. We showed that our framework outperforms two state-of-the-art baselines introduced by Pang and Lee (2005): one that ignores authorship information, and another that considers only the model learned for the author of the text whose polarity we want to infer. These results support our hypothesis that the way sentiment is expressed is often author-dependent, and shows that our approach successfully harnesses this dependency to improve polarity inference performance.

Topic-based representations of authors suggest a way of measuring similarity between authors based on their texts, which can be used by our polarity inference framework. Such measures are expected to capture authors' interests and aspects of their authorship style, which is indicative of demographic attributes and personality traits. We hypothesize that compact representation of authors using topic distributions would help handle the inherent noisiness of large data sets of user-generated texts without losing much information, as it did on the authorship attribution task.

To test this hypothesis, we experimented with a simple variant of our polarity inference framework, which infers the polarity rating of a sentiment-bearing text q written by author a according to a weighted average[2]

$$\frac{\sum_{a' \in \mathcal{N}_a} w_{aa'} \tilde{r}_{a'q}}{\sum_{a' \in \mathcal{N}_a} w_{aa'}} \tag{18}$$

where $\mathcal{N}_a$ is the set of neighbors of author a, $w_{aa'}$ is a non-negative similarity weight assigned to each neighbor a', and $\tilde{r}_{a'q}$ is the polarity inferred by the inferrer of a' for q (each inferrer is a support vector regression model trained on the labeled texts by a'). The neighborhood $\mathcal{N}_a$ is obtained for each author a by learning a threshold on the number of similar authors to consider. This is done by performing five-fold cross validation on a's set of labeled texts to find the threshold that minimizes the root mean squared error (RMSE) out of a set of candidate thresholds.

We compare the results obtained with baselines of equal weights (i.e., an unweighted average) and token frequency similarity with those obtained with similarity measures based on the AT and DADT topic models. The token frequency similarity measure is the cosine similarity of the frequency vectors of all the tokens in the authors' vocabularies. The AT and DADT similarity measures are calculated as one minus the Hellinger distance between the author topic distributions.

We ran this experiment on the IMDb62 data set. To test our approach in a variety of scenarios, we utilized the GivenX protocol, where each target author has exactly X training samples (Breese, Heckerman, and Kadie 1998). Specifically, we performed ten-fold cross validation over *authors*, where we partitioned the authors into ten folds and iterated over the folds, using nine folds as the training folds and the remaining fold as the test fold. The model was trained on all the labeled texts (IMDb62 reviews, each with a polarity rating assigned by its author) by the authors in the training folds, and exactly X labeled texts by each target author in the test fold. The model was then tested on the remaining samples by each target author. This process was repeated five times

2 This is not the strongest variant of those explored by Seroussi, Zukerman, and Bohnert (2010) and Seroussi (2012), where we found that normalizing the inferences from the neighborhood and considering a model trained on author a's texts improves performance. Using a simple weighted average allows us to compare similarity measures independently of these enhancements.

with different random seeds, and the RMSE was averaged across folds (here and in the next section we use a paired two-tailed t-test to measure statistical significance, as polarity inference and rating prediction are regression problems). Note that the GivenX protocol cannot be used to reliably compare the performance of the same method across different X values (e.g., testing how the performance of a method varies from Given1 to Given100), because the test samples vary across X values. Rather, we use this protocol to compare different methods under the same conditions, e.g., by comparing the performance of using DADT to that of employing equal weights under the Given10 scenario.

Figure 11 presents the results of this experiment. As the figure shows, the AT and DADT similarity measures outperformed the baselines and performed comparably to each other (the differences between either AT or DADT and the baselines are statistically significant in all cases except for Given5 and Given10 with the token frequency measure, whereas the differences between DADT and AT are not statistically significant in all cases). It is worth noting that we did not tune AT's and DADT's parameters. Instead, we used the settings that yielded the best authorship attribution performance on the IMDb62 data set (Section 5.4). It appears that in this case DADT's approach of de-noising the author representations by modeling authors and documents over two disjoint sets of topics is of little benefit in comparison with AT's approach of using only author topics. This may appear to stand in contrast to the results of our authorship attribution experiments (Section 5), but it could be because the similarity measures do not require the models' full discriminatory power, which is where DADT's strengths lie (Section 3.4.3). Nonetheless, we are encouraged by the fact that using either AT or DADT yielded better results than both the equal weights and token frequency baselines in most cases. This is despite the fact that these models operate in a space of lower dimensionality than the token frequency measure, which demonstrates the strength of topic-based approaches for author representation.

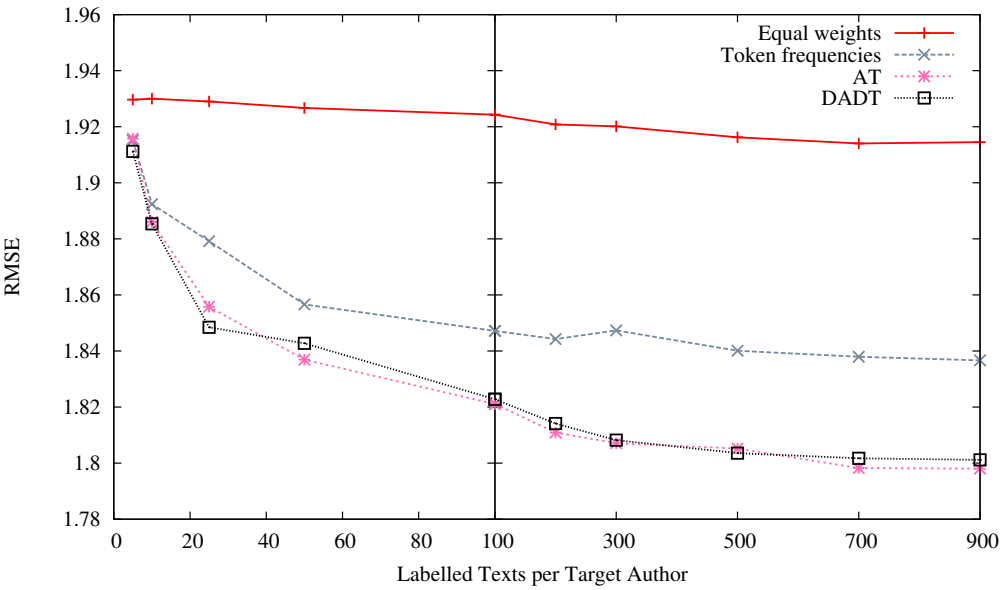


Figure 11
Author-aware polarity inference with topic models (data set: IMDb62).

6.3 Text-Aware Rating Prediction

Recommender systems help users deal with information overload by finding and recommending items of personal interest (Resnick and Varian 1997). **Rating prediction** is a core component of many recommender systems (Herlocker et al. 1999). Recently, rating prediction algorithms that are based on matrix factorization have become increasingly popular, due to their high accuracy and scalability (Koren, Bell, and Volinsky 2009). However, such algorithms often deliver inaccurate rating predictions for users with few ratings (this is known as the **new user problem**). We introduced an extension to the matrix factorization algorithm that considers user attributes when generating rating predictions (Seroussi, Bohnert, and Zukerman 2011; Seroussi 2012). We showed that using either demographic attributes or text-based attributes extracted with the LDA-S model, which is equivalent to AT (Section 3.2.3), outperforms state-of-the-art baselines that consider only ratings, thereby enabling more accurate generation of personalized rating predictions for new users. In the case of AT, these predictions are generated without requiring users to explicitly supply any information about themselves and their preferences.

Our framework predicts the rating user u would give to item i by switching between our attribute-based model and Koren, Bell, and Volinsky's (2009) ratings-only model according to an empirically set threshold n on the size of user u's known rating set $\mathcal{R}_u$:

$$\hat{r}_{ui} = \begin{cases} \mu + b_i^{(I)} + \sum_{t=1}^{T} p(t|u) \left(b_t^{(A)} + \boldsymbol{z}_{\cdot t}^{\top} \boldsymbol{y}_{\cdot i} \right) & |\mathcal{R}_u| < n \\ \mu + b_u^{(U)} + b_i^{(I)} + \boldsymbol{x}_{\cdot u}^{\top} \boldsymbol{y}_{\cdot i} & \text{otherwise} \end{cases} \tag{19}$$

where μ is the global rating mean; $b_u^{(U)}$, $b_i^{(I)}$, and $b_t^{(A)}$ are the user, item, and attribute biases, respectively; and $\boldsymbol{x}_{\cdot u}$, $\boldsymbol{y}_{\cdot i}$, and $\boldsymbol{z}_{\cdot t}$ denote the u-th, i-th, and t-th columns of the user, item, and attribute factor matrices $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$, respectively. The probability of a user u having one of the T attributes t is denoted by $p(t|u)$, which in the case of AT and DADT is the user's probability of using the author topic t, i.e., $\theta_{ut}^{(A)}$ (each topic model is inferred from the texts written by the user). We infer the biases and factor matrices using gradient descent in two stages (all the available ratings are used in both stages): (1) infer the ratings-only part of the model; and (2) infer the attribute-based part of the model, assuming that the item biases and factor matrix are given (Seroussi, Bohnert, and Zukerman 2011).

We ran Given0 and Given1 experiments on the IMDb1M data set, where the training set consisted of message board posts and rated reviews, and calculated the RMSE on the test ratings (the reviews associated with these ratings were hidden from the models). The baseline methods were non-personalized prediction ($\mu + b_i^{(I)}$, which is roughly equivalent to item i's rating mean), and the personalized, ratings-only model, which could only be used in the Given1 case. We set the number of author topics to 75 for Given0 and 125 for Given1, as this yielded the best results for AT (out of 5, 10, 25, 50, 75, 100, 125, and 150). In DADT's case, we used additional five document topics, which yielded the best results (out of 1, 5, 10, and 25). As the attribute-based model is sensitive only to the number of author topics, this enabled us to perform a fair comparison between the two models.

The results of this experiment are presented in Table 6. Both AT and DADT outperformed the baselines, which supports our hypothesis that considering user texts by using topic-based author representations can yield personalized and accurate rating

Table 6
Text-aware rating prediction with AT and DADT (data set: IMDb1M).

Method	Given0	Given1
Non-personalized	2.733	2.691
Personalized (only ratings)	—	2.734
Personalized (AT)	**2.719**	**2.668**
Personalized (DADT)	**2.719**	2.678

predictions, potentially leading to improved recommendations. The reason DADT did not outperform AT may be that DADT tends to yield user representations that help discriminate between texts by individual users (as shown in our authorship attribution experiments), but such representations are not as useful when utilized as attributes, because the attribute-based model requires a representation that captures commonalities between users.

7. Conclusion and Future Work

In this article, we extended and added detail to the work of Seroussi, Zukerman, and Bohnert (2011) and Seroussi, Bohnert, and Zukerman (2012) by reporting additional experimental results and applications of topic-based author representations that go beyond traditional authorship attribution. We provided experimental results for authorship attribution methods that are based on three topic models (LDA, AT, and DADT) for several scenarios where the number of authors varies from three to about 20,000. Specifically, we showed that in most cases, a probabilistic approach that is based on our DADT model (DADT-P) yields the best results, outperforming methods based on LDA and AT, as well as a Token SVM baseline. This indicates that our topic-based approach successfully captures indicators of authors' style (which is indicative of author characteristics such as demographic attributes and personality traits) as reflected by their texts. We harnessed this property when we used AT and DADT to uncover the authors of anonymous reviews where the training texts are multi-authored, improve performance when measuring similarity between authors based on their texts in our polarity inference framework, and obtain compact representations of users for our rating prediction framework.

The work presented in this article can be extended in many ways. One direction would be to address the limitation of relying only on unigrams as features by considering word order. This can possibly be pursued by adding author-awareness to Griffiths et al.'s (2004) HMM-LDA model, which considers word order by combining LDA with a Hidden Markov Model. Author awareness can also be introduced into the models suggested by Wallach (2006) and Wang, McCallum, and Wei (2007), who made each word dependent on both its topic and on the previous word (at a considerable computational cost). A more general alternative would be to enable the use of various feature types, for example, by incorporating conditional random fields into DADT in a manner similar to Zhu and Xing's (2010) model. This direction can also be pursued by using DADT-P in an ensemble with SVMs that can be trained on feature types other than token unigrams, which may also have the added value of combining the strengths of DADT with those of the SVM approach (Section 5.4). Testing these approaches with character *n*-grams would be of particular interest, as they often deliver strong performance, sometimes outperforming token unigrams (Koppel, Schler, and Argamon 2009).

Our DADT-P method can be extended to handle situations where the test texts may have not been written by any of the candidate authors (i.e., open-set attribution and verification, described in Section 2). A fairly straightforward approach consists of setting a threshold on the probability assigned to the selected author based on performance on held-out data—if the probability of the selected author is below the threshold, then "unknown author" is returned. This approach was successfully used by Tanguy et al. (2011) in conjunction with a maximum entropy classifier.

Another potential extension would be to automatically infer the optimal number of author and document topics. This is likely to yield improved results, because the number of topics had the largest impact on performance among the parameters considered in our experiments (Section 5.5). In addition, our models can be extended to address semi-supervised authorship attribution, and may potentially be applied to any scenario where user-generated texts are available, going beyond the applications presented in Section 6.

Acknowledgments

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 Judgment data set, Mark Carman for fruitful discussions on topic modeling, and the anonymous reviewers for their insightful comments.

References

Argamon, Shlomo 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.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

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

Blei, David M. and Jon D. McAuliffe. 2007. Supervised topic models. In *NIPS 2007: Proceedings of the 21st Annual Conference on Neural Information Processing Systems*, pages 121–128, Vancouver.

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

Breese, John S., David Heckerman, and Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *UAI 1998: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52, Madison, WI.

Brennan, Michael and Rachel Greenstadt. 2009. Practical attacks against authorship recognition techniques. In *IAAI 2009: Proceedings of the 21st Conference on Innovative Applications of Artificial Intelligence*, pages 60–65, Pasadena, CA.

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

Fan, Rong-En, 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*, 9:1871–1874.

Fog, Agner. 2008. Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics, Simulation and Computation*, 37(2):258–273.

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

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *NIPS 2004: Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, pages 537–544, Vancouver.

Groves, Trish. 2010. Is open peer review the fairest system? Yes. *BMJ*, 341:c6424.

Herlocker, Jonathan L., Joseph A. Konstan, Al Borchers, and John Riedl. 1999. An algorithmic framework for performing collaborative filtering. In *SIGIR 1999: Proceedings of the 22nd International ACM SIGIR Conference on Research and*

Development in Information Retrieval, pages 230–237, Berkeley, CA.

Juola, Patrick. 2004. Ad-hoc authorship attribution competition. In *ALLC-ACH 2004: Proceedings of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities*, pages 175–176, Göteborg.

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

Kacmarcik, Gary and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *COLING-ACL 2006: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (Main Conference Poster Sessions)*, pages 444–451, Sydney.

Kern, Roman, 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.

Koppel, Moshe and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *ICML 2004: Proceedings of the 21st International Conference on Machine Learning*, pages 62–68, Banff.

Koppel, Moshe, 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.

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

Koren, Yehuda, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37.

Kourtis, Ioannis 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.

Lacoste-Julien, Simon, 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.

Liu, Bing and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*. Springer US, pages 415–463.

Luyckx, Kim and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1):35–55.

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

Mimno, David 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.

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

Nanavati, Mihir, Nathan Taylor, William Aiello, and Andrew Warfield. 2011. Herbert West—deanonymizer. In *HotSec'11: Proceedings of the 6th USENIX Workshop on Hot Topics in Security*, San Francisco, CA.

Ng, Andrew Y. and Michael I. Jordan. 2001. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *NIPS 2001: Proceedings of the 15th Annual Conference on Neural Information Processing Systems*, pages 841–848, Vancouver.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL 2005: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, MI.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pearl, Lisa and Mark Steyvers. 2012. Detecting authorship deception: A supervised machine learning approach using author writeprints. *Literary and Linguistic Computing*, 27(2):183–196.

Rajkumar, Arun, 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.

Ramage, Daniel, 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.

Resnick, Paul and Hal R. Varian. 1997. Recommender systems. *Communications of the ACM*, 40(3):56–58.

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

Rosen-Zvi, Michal, 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.

Rosen-Zvi, Michal, 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.

Salton, Gerard. 1971. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ.

Salton, Gerard. 1981. A blueprint for automatic indexing. *SIGIR Forum*, 16(2):22–38.

Sanderson, Conrad and Simon Guenter. 2006. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In *EMNLP 2006: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491, Sydney.

Schler, Jonathan, 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.

Seroussi, Yanir. 2012. *Text Mining and Rating Prediction with Topical User Models*. Ph.D. thesis, Faculty of Information Technology, Monash University, Clayton, Victoria, Australia.

Seroussi, Yanir, Fabian Bohnert, and Ingrid Zukerman. 2011. Personalized rating prediction for new users using latent factor models. In *HT 2011: Proceedings of the 22nd International ACM Conference on Hypertext and Hypermedia*, pages 47–56, Eindhoven.

Seroussi, Yanir, Fabian Bohnert, and Ingrid Zukerman. 2012. Authorship attribution with author-aware topic models. In *ACL 2012: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–269, Jeju Island.

Seroussi, Yanir, Russell Smyth, and Ingrid Zukerman. 2011. Ghosts from the High Court's past: Evidence from computational linguistics for Dixon ghosting for McTiernan and Rich. *University of New South Wales Law Journal*, 34(3):984–1005.

Seroussi, Yanir, 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.

Seroussi, Yanir, 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.

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

Steyvers, Mark 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*. Lawrence Erlbaum Associates, pages 427–448.

Tanguy, Ludovic, 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.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006.

Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Wallach, Hanna M. 2006. Topic modeling: Beyond bag-of-words. In *ICML 2006: Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984, Pittsburgh, PA.

Wallach, Hanna M., 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 1,973–1,981, Vancouver.

Wang, Xuerui, Andrew McCallum, and Xing Wei. 2007. Topical N-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM 2007: Proceedings of the 7th IEEE International Conference on Data Mining*, pages 697–702, Omaha, NE.

Webb, Geoffrey I., Janice R. Boughton, and Zhihai Wang. 2005. Not so naive Bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24.

Wong, Sze-Meng Jojo, Mark Dras, and Mark Johnson. 2011. Topic modeling for native language identification. In *ALTA 2011: Proceedings of the Australasian Language Technology Association Workshop*, pages 115–124, Canberra.

Zhu, Jun, Amr Ahmed, and Eric P. Xing. 2009. MedLDA: Maximum margin supervised topic models for regression and classification. In *ICML 2009: Proceedings of the 26th International Conference on Machine Learning*, pages 1,257–1,264, Montreal.

Zhu, Jun and Eric P. Xing. 2010. Conditional topic random fields. In *ICML 2010: Proceedings of the 27th International Conference on Machine Learning*, pages 1,239–1,246, Haifa.