Digital Scholarship in the Humanities Advance Access published June 23, 2015

Profile-Based Authorship Analysis

Jonathan Dunn, Shlomo Argamon, Amin Rasooli and Geet Kumar
Illinois Institute of Technology, Chicago, IL, USA

Abstract

This article presents a profile-based authorship analysis method which first categorizes texts according to social and conceptual characteristics of their author (e.g. Sex and Political Ideology) and then combines these profiles for two authorship analysis tasks: (1) determining shared authorship of pairs of texts without a set of candidate authors and (2) clustering texts according to characteristics of their authors in order to provide an analysis of the types of individuals represented in the data set. The first task outperforms Burrows' Delta by a wide margin on short texts and a small margin on long texts. The second task has no such benchmark with existing methods. The data set for evaluating the method consists of speeches from the US House and Senate from 1995 to 2013. This data set contains both a large number of texts (42,000 in the test sets) and a large number of speakers (over 800). The article shows that this approach to authorship analysis is more accurate than existing approaches given a data set with hundreds of authors. Further, this profile-based method makes new types of analysis possible by looking at types of individuals as well as at specific individuals.

Correspondence:
Jonathan Dunn, Illinois Institute of Technology
SB 237C
10 W. 31st Street
Chicago, IL 60616, USA.
E-mail: jonathan.edwin.dunn@gmail.com

1 Introduction

Authorship analysis has traditionally been concerned with two tasks: first, distinguishing texts written by different individuals (Hoover, 2004; Abbasi and Chen, 2008; Koppel *et al.*, 2011, 2012); second, distinguishing texts written by individuals from different social groups (Koppel *et al.*, 2003; Yu *et al.*, 2008; Garera and Yarowsky, 2009; Mukherjee and Liu, 2010; Nguyen *et al.*, 2011; Sarawgi *et al.*, 2011). The present work combines both tasks, identifying texts as written by members of larger groups as an intermediary step in distinguishing texts written by different individuals.

This ensemble approach presents, on the one hand, an additional layer of computation for authorship analysis over direct approaches. On the other hand, though, it also presents a number of distinct advantages. First, it remains efficient as the set of speakers grows. For example, the data set used for evaluation consists of over 800 speakers, or over 300,000 direct comparisons between authors (Luyckx and Daelemans, 2010). A profile-based method, however, sees no increase in comparisons as the number of speakers grows. Second, it allows the identification of the dimension of difference between the authors of two texts. This is important given the large number of similar individuals in the world. In other words, in the real world there are very large numbers of individuals of the same age and sex, with the same educational and linguistic background, from the same geographic area. A direct comparison of texts written by individuals who differed only by age, then, would output only that the texts are very similar. The profile-based method presented here is able to indicate precisely the profile which distinguishes such similar individuals. Third, although this is not undertaken in the present work, a profile-based method is capable of cross-genre comparisons by building profiles within

doi:10.1093/llc/fqv019

specific genres and testing authorship using these profiles.

This profile-based method is used here for two tasks. The first is the fundamental task of authorship analysis, taking a pair of texts and deciding whether or not they were produced by the same individual without a set of candidate authors. The second task is to characterize the authors of texts in a given data set by clustering texts together according to the characteristics of the author. This provides a way of viewing the data set that focuses on types of individuals rather than on specific individuals. This is an important task because of the very large number of similar individuals in any large and representative data set.

2 Building Author Profiles

The first step is to build profiles of a text's author. This means determining, given the text, the characteristics of the individual who produced that text. The method discussed here uses seven social characteristics (Age, Geographic Location, Previous Military Service, Race, Religion, and Sex) and four conceptual characteristics (Party Membership, two dimensions of Special Interest Group Ratings, and Chamber Membership). These author characteristics will be discussed further in Section 2.3.

The profile building process has three important parts. First, the textual features used are discussed in Section 2.1; second, the supervised learning algorithm used is discussed in Section 2.2; third, the data and meta-data used for evaluating the method are discussed in Section 2.3.

2.1 Features

Traditional authorship studies use word-form and part-of-speech n-grams (Stamatatos, 2009; Grieve, 2007, for a range of feature types). While these features have been shown to be effective, there is a constant trade-off required between several factors (Antonia *et al.*, 2013). Shorter n-grams (e.g. unigrams) can be less sparse than longer n-grams, but also miss phrases and idioms which can be important for authorship analysis. Word-forms are sparser than part-of-speech (POS) tags, but also contain a wider range of information, although that information is often topic-dependent. Choosing only a single level of representation increases efficiency by reducing the number of features used, but does so at a cost to performance. On the other hand, adding multiple n-gram lengths can improve text representation, but does so at the expense of efficiency, creating vectors that are very large and very sparse.

Table 1 Feature extraction and evaluation algorithm

1	Let F be the frequency threshold for retaining feature f_i
2	Let T be a list of representation types (e.g. pos, word-form)
3	Let N be a range of n-gram lengths (e.g. 1–4)
4	For n_i in N:
5	For t_j in T:
6	Extract all features of type t_j and length n_i
7	Count all extracted features
8	Remove all features with frequency below F

The method used here combines n-grams of multiple lengths, using both word-form and part-of-speech representations (which have shown to be useful; Hirst and Feiguina, 2007), without sacrificing efficiency. The method works by extracting each type of feature (e.g. word-form unigrams), evaluating the instances of this feature for their usefulness, and discarding features below a given threshold before continuing the feature extraction process. Table 1 shows the algorithm for extracting and evaluating features. This algorithm keeps the most frequent features for each n-gram range and representation type, remaining efficient by evaluating and removing candidates for each combination of n-gram range and feature type which fall below a set frequency threshold. This produces a feature set similar to traditional features, but allows the information from a range of different combinations (e.g. word-form unigrams, part-of-speech trigrams, etc.) to be captured without creating large sparse vectors. Features are calculated using their relative frequency in each text. Part-of-speech tags are used to disambiguate word-forms so that, for example, different uses of 'have' are computed separately.

For the purposes of this study, the n-gram range was set to 1 through 3, with a frequency threshold of 2.5% of documents (e.g. a feature must occur in at least 2.5% of documents in order to be selected). This produced a total of 6,618 features, with the

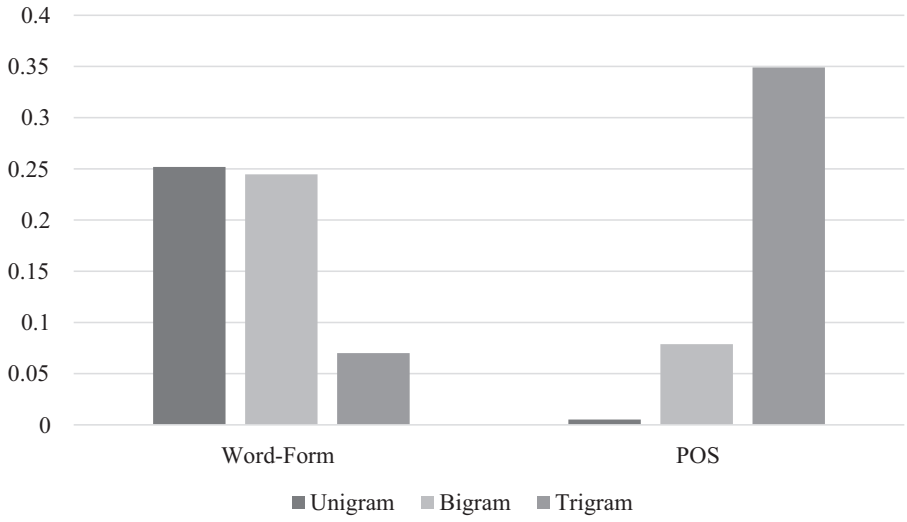

Fig. 1 Distribution of features across representation type and n-gram range, by percentage

distribution as shown in Fig. 1. Two interesting observations can be made from this distribution. First, the two types of representation differ in the usefulness of longer context windows: part-of-speech tags become more useful in longer sequences while word-forms become less useful (by 'useful' we mean less sparse; a more precise way of putting this is 'potentially useful'). Part of this is a result of the limited set of part-of-speech tags. But it also shows that the two types of representations work best at opposite context windows. Second, word-form unigrams, the most commonly used feature, constitute the largest single group of useful features. However, if we were limited to only this group we would be missing 75% of the total potentially useful features. This shows the practical benefit of this type of feature extraction, to include the largest number of useful features while at the same time reducing the sparsity of vectors by evaluating each iteration of n-gram length and representation type before continuing.

2.2 Algorithms

Given this set of textual features, we use supervised learning algorithms to build author profiles. While unsupervised algorithms do not require annotated data, they are unlikely to move beyond the primary dimensions of variation. In other words, much of the variation produced by this feature set will be topic-dependent, but topic-dependent differences are not necessarily the sort of variations that we need to build all of the profiles we are interested in.

We build profiles using the same feature set for each author characteristic, choosing not to optimize features for individual profiles. This allows the comparison of predictive features across profiles in order to determine how mutually dependent the profiles are. Given the feature extraction algorithm, however, this single feature set includes the best features of many types of features. A linear support vector machine was used, the sequential minimal optimization algorithm (Platt, 1998; Keerthi *et al.*, 2001), available through Weka (Hall *et al.*, 2009). Support Vector Machine (SVM) because it combines the high performance of SVMs with the ability to examine individual feature weights, an important criterion given the need to maintain independent profiles. While there are many other options (Jockers and Witten, 2010), the linear SVM is used because it provides these important capabilities.

A development corpus was used to optimize SVM parameters and features. This development corpus used the 104th and 105th congresses as training data and the 107th as testing data. Because this development corpus overlaps to some degree with the full corpus used for testing the author falsification system, a separate corpus containing speeches from the 110th through 112th congresses was used

as a held-out test set for both profiling and falsification, as discussed in Section 5.1. The parameter C was set at 0.01 for all profiles based on this development corpus. Vectors were normalized during training. Tests of individual feature types (e.g. word-form unigrams) showed that many feature types increased performance, so that the final inclusive feature set was used.

2.3 Data set

The textual data consists of individual speeches given in the US House and Senate as recorded in the *Congressional Record* (Gentzkow and Shapiro, 2013) from the 104th congress (beginning in 1995) through the 109th congress (ending in 2006). This data set was divided into a training set (104th through 108th congresses) and a testing set (109th congress). Separate training and testing sets were used instead of cross-validation in order to limit the influence of debate-specific topic-dependent features on profile-building (e.g. there is no chance that speeches from the same debate occur in both training and testing data this way). In order to minimize differences between the chambers, chamber-specific procedural content was removed (e.g. 'Mr. Speaker' and 'Mr. President'), although some less common differences remain. This training/testing division was used to test different components of the system (e.g. to optimize SVM parameters, to optimize the authorship falsification parameters, etc.), and the entire system is then tested on a held-out test set consisting of speeches from the 110th through 112th congresses (2007–13).

Two versions of this data set were used: one containing all speeches longer than 500 words (the text limit used to trim procedural and formulaic speeches) and one containing a single aggregated text for each speaker in each congress (increasing the length of each text). These two data sets represent a trade-off between a larger number of smaller texts and a smaller number of longer texts. Descriptive statistics for these different versions of the data set are shown in Table 2. The number of instances for the version with individual speeches is much higher (85,506 versus 2,665 for the training set), while the average length of each text is much shorter (1,123 versus 49,020 words).

A total of eleven author characteristics were included in the meta-data. Some of these are social characteristics, traditional to sociolinguistic studies of language variation, and some of these are conceptual characteristics related to political ideology that are somewhat new to studies of language variation (Laver *et al.*, 2003; Thomas *et al.*, 2006; Diermeier *et al.*, 2011; Dahllof, 2012; Iyyer *et al.*, 2014). The idea behind the conceptual characteristics is that the speakers in this data set form speech communities based on shared ideologies and that these speech communities engage in an internal dialogue and partake of the same media sources. In other words, these communities have the same properties as traditional speech communities and we would expect that at least some linguistic variations have been selected by one community but not by another.

For the purposes of evaluation, all scalar attributes were reduced to categorical attributes, and the number of categories for each attribute was kept to a minimum in order to maintain robust classes. For example, age is a scalar variable representing how old the speakers are, but we do not expect significant differences from one year to another; rather, we expect generational changes. Thus, rather than treat age as a scalar attribute (i.e. as a number), we treat it as a categorical attribute so that speakers of similar ages are combined into generations. Classification is then performed to try to distinguish between generations. This means that for many profiles a decision had to be made about where to place the cutting line between two classes (e.g. the division between generations). For scalar attributes, this was done using

Table 2 Descriptive statistics for each version of the data set

Speeches	Period	Instances	Average words	Range of words
Aggregated	Train: 104–108	2,665	49,020	639–545,828
Aggregated	Test: 109	523	46,509	1,303–538,195
Individual	Train: 104–108	85,506	1,123	500–6,335
Individual	Test: 109	15,694	1,131	500–6,234

equal-width binning to divide the attribute space into two sections. Equal-width binning divides scalar attributes into categorical attributes, with each category representing the same span. For example, if the binning is done by years then each category will contain the same number of years. For nominal attributes, this was done manually by deciding how to best categorize individuals in a way that balances accuracy with robust classes (e.g. for religion, there are many reported values such as Baptist and Southern Baptist that overlap; classes are reduced to contain the smallest number of meaningful categories). This reduction was done on the original meta-data obtained from the CQ Congress Database and the Votesmart.org database in order to produce a small number of intuitively coherent labels. No accuracy testing was used to choose a particular set of labels.

The social characteristics of speakers include Sex, Age, Religion (reported), Race, Geographic Location (of office), and Previous Military Experience. Age was divided into two classes: those born before 1938 and those born after. Race was reduced to a binary distinction between white and non-white speakers, given the underrepresentation of minorities in this data set. Religion was divided into Catholic Christians, Non-Catholic Christians, and Non-Christians. While not ideal (e.g. Jews and Atheists are lumped into a single category), this was chosen given the large majority of speakers reporting Christian affiliations (leaving the remaining category as a catch-all). Previous Military Service was likewise reduced to a binary distinction, with any sort of service from National Guard to Active Duty constituting a single category. Sex was reduced to a strict Female/Male distinction, again appropriate given the individuals contained in this data set. Geographic Location, of the office the individual holds rather than birthplace, is divided into North, South, Midwest, and West.

Some social characteristics that can influence linguistic variations were not included. For example, level of education was not included, largely because most members of congress fall into the same category. Economic information (e.g. net worth of members of congress) was also not included, in part because of the skewed representation and in part because of lack of consistent information. For example, all members of congress, by virtue of their salary alone, will be upper-middle-class or higher. What might be more important is the economic status of the speaker before becoming a member of congress (e.g. someone born wealthy versus someone born into poverty). However, this information is not consistently available for speakers. Related to this, we use as geographic location the state in which the speaker holds office. But many people hold office far from where they were born and/or raised. This again is a difficulty of gathering consistently reliable information: first, this information is not available for many non-famous speakers; second, there is the difficulty of setting firm guidelines for each category: how old did the speaker have to be to consider the location of birth different from the location of being raised? For these reasons, these social characteristics, which may have yielded useful information, are not used in the current study.

The conceptual (i.e. ideological) characteristics of speakers include Party Membership, Chamber Membership, and two dimensions of special interest group ratings. Party Membership is a nominal distinction between Republicans and Democrats, and Chamber Membership is a nominal distinction between the House and the Senate. The two chambers have different compositions, the Senate being more elite than the House. This distinction is complicated by the fact that the two bodies have different procedures (e.g. 'unanimous consent' is unique to the senate) and different topics of debate. Thus, because we train and test on both Chambers together, it is likely that some of the features separating the Chambers are artifacts of this data set while others are legitimate differences in language use.

A problematic issue results when some aspect of a speaker's characteristics change over the course of time represented by the data set. For example, a speaker may change parties, convert to a new religion, or move between chambers. First, chamber information is taken from the meta-data of individual speeches; thus, changes between chambers are accounted for and each speech is categorized according to the chamber in which it was given. Second, other changes in speaker meta-data are not taken into account. However, this does not pose a significant challenge. For example, no

speakers changed in gender or race during this data set. Change in political party is rare but does occur. For example, in the time period covered four senators changes their political party. Two of these changes occurred at the very beginning (Ben Campbell) and very end (Joe Lieberman) of the data set, and thus did not cause changes within the data set (e.g. they were accurately represented). One of these changes (Robert Smith) lasted only a few months, thus causing limited difficulty. Only one of these changes (Jim Jeffords) is meaningful in the data set. Jeffords changes from a Republican affiliation to an independent affiliation, and in this way still did not alter the Republican versus Democrat classification. Another possible complication is that a speaker may adopt, for rhetorical purposes, a position that speaker does not hold. Whether this phenomenon is significant enough to interfere with the classification of political stance is an empirical question.

Two dimensions of special interest group ratings (e.g. the League of Conservation Voters) were also used. The analysis started with 134 interest group ratings, some with relatively sparse coverage across speakers, taken from the CQ Congress database and the Votesmart.org database. The first task was to reduce these ratings into a small number of robust and interpretable groups that represent a single dimension of the speaker's ideology. Principal Components Analysis was used to group interest ratings together. Missing values were replaced with the scale's mean value. The varimax rotation method was used with Kaiser normalization in order to create components with unique memberships. Thirteen components were identified with eigenvalues above 1 before rotation.

Interest group ratings are included in a given component if (1) the component loading is above 0.400, (2) the component loading is within 0.150 of the highest member of the component, and (3) the component loading is at least 0.300 above the same rating's loading in a different component. Only ten components had at least one member according to these criteria. To ensure adequate representation, we only included components in which no more than 100 speakers were not rated by any of the groups in the components. This last condition eliminated all but the first three components; the other components had relatively sparse coverage and were likely influenced by the use of mean values to replace missing instances. The three components, shown in Table 3 together with the variance explained by each, represent the relation between the Government and (1) Institutions (for example, labor unions and universities and businesses), (2) Humans (for example, immigrants and children and the needy, regardless of nationality), and (3) Animals. Together, these components account for 67.69% of the variation in interest group ratings. Each of these components is turned into a single score by taking the average of all member ratings. These Special Interest Group components are indicated with the abbreviation SIG.

Table 3 Ideology components with percent of variance explained

Numbers	Name	Scale	Variance
1	Government and Institutions	100 = Liberal, 0 = Conservative	37.08%
2	Government and Humans	100 = Liberal, 0 = Conservative	26.48%
3	Government and Animals	100 = Liberal, 0 = Conservative	4.13%

Only two of these dimensions of special interest group ratings are used because of the high correlation between the Government and Institutions dimension and the Government and Humans dimension. Another way of measuring political ideology is through roll call voting (Poole and Rosenthal, 2007). We cannot use roll call voting measures across Chambers because the measure is not directly comparable across individual voting bodies. We can, however, examine the correlation between our chamber-independent special interest group measures and the roll call voting scores within each chamber. Table 4 shows the Pearson correlations between the special interest group rating components and the roll call voting measures (called Dw-Nominate; there are two dimensions of this measure in the US context, as indicated by the row headings).

This table shows us two important things: first, the Government and Institutions dimension and the

Table 4 Pearson's correlations between ideology measures, House only/Senate only

	DW-N.1	DW-N.2	SIG: Inst.	SIG: Humans	SIG: Animal
DW-Nominate 1	1	0.055/0.230	0.954/0.958	0.887/0.938	0.689/0.758
DW-Nominate 2	0.055/0.230	1	0.017/0.284	0.049/0.224	0.381/0.019
SIG: Institutions	0.954/0.958	0.017/0.284	1	0.887/0.951	0.751/0.782
SIG: Humans	0.887/0.938	0.049/0.224	0.887/0.951	1	0.635/0.776
SIG: Animals	0.689/0.758	0.381/0.019	0.751/0.782	0.635/0.776	1

Table 5 Top five interest groups in component 1, government and institutions

Component	Interest group	Loading
1	Committee on Political Education of the AFLCIO	0.873
1	United Auto Workers	0.869
1	American Federation of State County Municipal Employees	0.864
1	Communication Workers of America	0.862
1	American Federation of Government Employees	0.862

Table 6 Top five interest groups in component 3, government and animals

Component	Interest group	Loading
3	Animal Welfare Institute	0.695
3	Humane Society of the US	0.593
3	Born Free USA	0.560
3	American Humane Association	0.560
3	American Society for the Prevention of Cruelty to Animals	0.560

Government and Animals dimension are correlated (0.751/0.782) but not as highly as the first with the Government and Humans dimension (0.887/0.951). Thus, we discard that dimension in order to reduce the number of related profiles, but keep the Government and Animals component because it contains unique information. Second, this table shows us that the special interest group ratings are highly correlated with the roll call voting based measure (0.954/0.958, the sign is irrelevant). Thus, even though we cannot use the roll call voting measure for profile-building across chambers, we can use it to show that our special interest group ratings are valid measures of political ideology.

The first component, concerning the relation between Government and Institutions, contains twenty-three interest groups, some of which are shown in Table 5. First, this component has a large number of unions (eleven, or 48% of the component). This reflects a question of how involved the government should be in regulating and monitoring the relationships between businesses and industries, on the one hand, and individual workers, on the other hand. A higher score indicates agreement with a more active role for government. Second, this component has a number of groups concerned with the rights of minorities or less powerful members of society in relation to institutions, both public and private. Thus, the component contains interest groups seeking government involvement and protection of African Americans, Children, and the Elderly (six groups, or 26%). Again, a higher score indicates agreement with a more active role for government in mediating the relationships between institutions and individuals.

The second component, concerning the relation between Government and Animals, contains six interest groups, some of which are shown in Table 6. The focus of this component is on the role of government in representing and producing policy in respect to nonhumans. A higher score indicates agreement that government should have an active role in nonhuman life.

3 Profile Building Results

The first step in profile-based authorship analysis is to profile individual texts according to characteristics of the text's author, in this case using a supervised approach with texts labeled with author characteristics for training models. Profile models were trained on the 104th through 108th congresses

with the feature set described in Section 2.1 using a linear SVM classifier (this overlaps somewhat with the development corpus used to optimize SVM parameter settings and the authorship falsification settings; thus, both profile-building and authorship falsification are evaluated using a held-out test set in Section 5.1.). Three of the classes were strongly imbalanced: Sex (86.42% majority baseline in the testing set), Race (84.89% baseline), and Religion (55.35% baseline, with three classes). These classes were balanced in the training set by randomly removing instances of the majority classes. The testing set was left unbalanced. Two classes changed over time, so that they were balanced in the training set but imbalanced in the testing set: Age (82.79% baseline in testing set) and Military Service (76.67% baseline in testing set). These profiles were left as-is in both training and testing sets. Profiles were modeled on individual speeches and on aggregated speeches (one text per individual per congress). The baseline results shown for each profile use the same linear SVM classifier but with only word-form unigram features.

We begin with the balanced (in the training set) social characteristics: Age, Previous Military Service, and Geographic Location (of office), shown in Fig. 2. For these balanced classes we look at three measures of profile-building performance: Precision (True Positives/True and False Positives), Recall (True Positives/True Positives and False Negatives), and F-Measure (a combination of Precision and Recall). Each of these measures is averaged across classes, weighted by the number of instances in each class.

Age improves with longer text lengths, from 0.669 to 0.757 F-Measure. At the same time, the largest area of improvement is in Precision, from 0.592 to 0.860. Classification by age does not improve above the baseline for the individual speeches data set, but improves significantly above the baseline in the aggregated data set. Classification by previous military service again shows an improvement with longer texts, albeit a smaller one, from 0.648 to 0.667 F-Measure. Here, again, the biggest improvement is with Precision, from 0.568 to 0.628. In this case there is no significant improvement over the word-form unigram baseline. Geographic Location shows the opposite trend, with performance actually declining with longer texts, showing that in at least one case a larger number of instances is better than longer texts per instance. At the same time, since Geographic Location has four classes, the profiling task is more difficult than for the other author characteristics. We see from these results that performance on balanced social characteristics is good on short texts and, generally, even better on aggregated

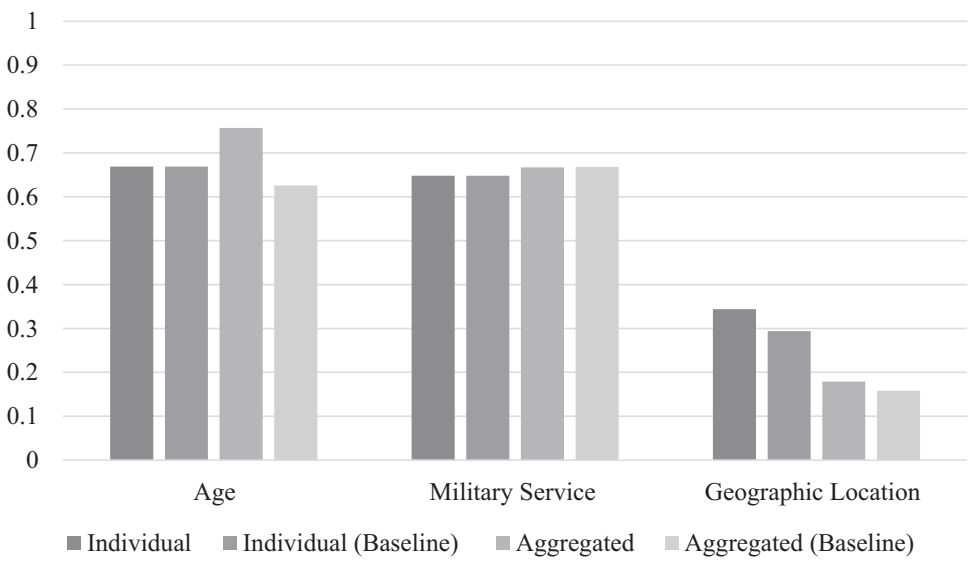

Fig. 2 Balanced social profiles: age, military service, geographic location by F-measure

texts. More importantly, we will see later that even profiles with less than ideal performance are useful for authorship analysis.

Both Age and Previous Military Service are balanced in the earlier training data and increasingly unbalanced in the testing data. One difficulty with age is that over the course of the data set the population changes significantly. For any number of bins, the majority class changes across training and testing data. At the same time, it is important to use a generational definition of age rather than the age of an individual at the time a speech is given. The same is true for Military Service, where the percentage of members of congress who have served in the military diminishes over time, no doubt related to the percentage of military service in the population as a whole for the represented generations. In the end, the performance of these dimensions is not far above the majority baseline by the testing set, given that they have become unbalanced. At the same time, however, the training set is balanced on its own (e.g. not forcibly balanced), so that these results are not simply inheriting the correct balance between the classes in order to improve performance. Thus, the performance on these profiles is meaningful, in spite of its unbalanced nature over time. This is not the case, however, with the remaining social profiles.

Next we look at the unbalanced social classes: Sex, Race, and Religion, shown in Table 7. These results are reported by accuracy and shown with a majority baseline (e.g. the accuracy achieved by simply choosing the majority class for each instance). All three characteristics remain below the majority baseline for both text lengths, not surprising given how large the majority class is. Further, because the classifier is trained on balanced data, the results are far below baseline. We also show the accuracy for models that were trained on unbalanced data, in the rightmost column of Table 7. For the models trained on unbalanced data, the accuracy hovers very slightly above the majority baseline. This is not a positive result, however, but only an artifact of the classifier reproducing the split between the two classes. In other words, classification by these attributes is not successful given this skewed data set.

Table 7 Unbalanced social characteristics: sex, race, religion

Class	Text length	Majority baseline (%)	Accuracy, bal. (%)	Accuracy, unbal. (%)
Sex	Individual	84.89	61.68	85.33
	Aggregated	85.25	60.27	85.05
Race	Individual	88.71	69.98	88.80
	Aggregated	86.42	74.90	86.59
Religion	Individual	52.51	40.32	52.56
	Aggregated	55.25	49.05	55.36

No baseline for word-form unigrams is shown for these classes because they fall below the majority baseline. The poor performance of these profiles is contrary to other studies (Koppel *et al.*, 2003; Argamon *et al.*, 2009a,b; Mukherjee and Liu, 2010; Yu, 2013). Only Yu dealt with a similar data set, and in that study the senate was not included because of underrepresentation of women; further, no predictions were made about the gender of authors of individual speeches. Thus, this finding of poor performance is not unexpected on this particular data set and does not license wider conclusions about these characteristics in other data sets.

Next we look at the conceptual characteristics that represent political ideology: Party Membership, Government and Institutions, Government and Animals, and Chamber Membership (which is partially an author characteristic and partially a genre characteristic), shown in Fig. 3. Each of the characteristics sees a significant performance boost with longer texts, and each performs well under both conditions. Simple party membership goes from 0.635 F-Measure with individual speeches to 0.791 with aggregated speeches; meaningful gains are also made above the word-form unigram baseline. Chamber membership goes from 0.869 F-Measure with individual speeches to 0.998 with aggregated speeches, with no improvement above the word-form unigram baseline for aggregated speeches but a small increase for individual speeches. This effect for chamber membership is related to genre differences rather than author differences. At the level of individual speeches, the classification is likely mostly by author characteristics, given that individual speeches are shorter and less likely to contain

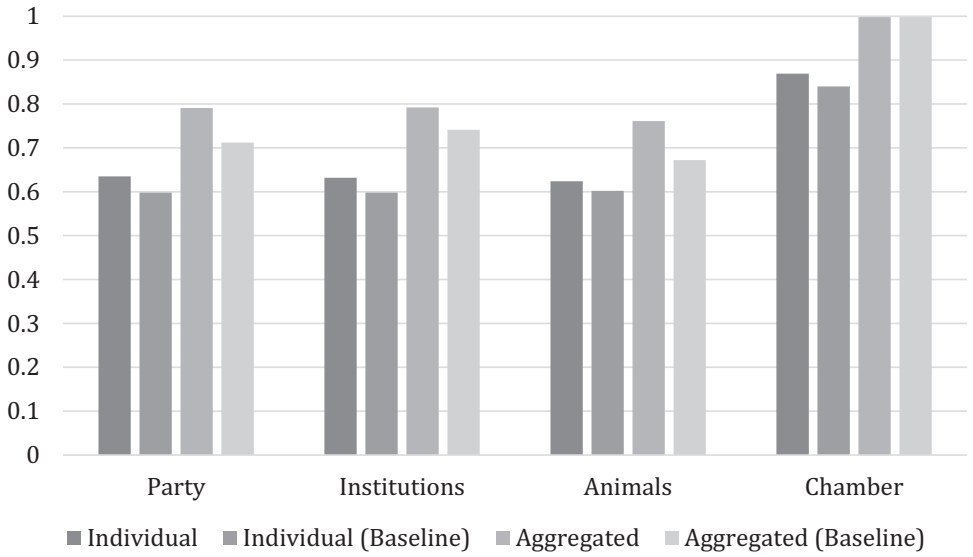

Fig. 3 Ideological characteristics: party, special interest groups, chamber: full [baseline]

chamber-specific procedural material. With aggregated speeches, however, classification is nearly perfect. This is not surprising because, at some point over a 2-year period, individuals will make some chamber-specific procedural statements.

The two components of special interest group ratings, here abbreviated as Institutions and Animals, perform at 0.792 (aggregated) and 0.761 (aggregated) F-Measures, respectively. This is good performance, and improves upon the word-form unigram baseline. These scalar measures were divided into two classes using equal-width binning. However, the measures are not normally distributed, with the majority of individuals falling to the extreme sides. Thus, there is a considerable overlap between these scales and party membership. We do not necessarily want to multiply the number of profiles which are capturing the same information as that would give extra weight to a single dimension. We test the independence of the profiles, among other things, in Section 4. For now, however, we test the relatedness of these three characteristics (Party Membership, Government and Institutions, Government and Animals) by viewing the three measures as independent raters trying to place the aggregated texts from the training corpus into a two-way categorization method (e.g. taking the actual profile values as each characteristic's rating). Given this scenario, we calculate inter-rater agreement between these measures using Fleiss' Kappa ($n = 2{,}576$; instances with missing values are removed). The agreement measured in this way is 0.732, showing a high level of consistency across the measures. We investigate in the next section how this influences the feature weights for each profile.

This leaves us with seven well-performing profiles of author characteristics, three of which often overlap. The real test for independence, however, is the degree to which the predictive textual features for each characteristic are correlated between profiles.

3.1 Profile-based author clustering

This profile-based method allows us to cluster texts together according to similarities of the texts' authors. In other words, instead of asking if the same individual wrote a text, we can ask whether the same type of individual wrote a text. Further, we can label the type of individuals using profiles in order to aid interpretation of the produced clusters. This produces a summary of the characteristics of the sorts of authors present in the data set. In cases with large

numbers of authors, especially when these authors are similar in many ways, this provides an introductory authorship analysis before the investigation of specific authorship. This analysis is useful, also, in cases where the likelihood of capturing specific authorship with confidence is somewhat small. Thus, this sort of analysis is meant more to proceed specific authorship identification and to operate when specific authorship identification is not practical, rather than to replace it.

In this section we first build clusters using the actual author characteristics of each text in the testing set. We then apply the same algorithms to the predicted author characteristics of the texts and evaluate the clusters according to how closely the textually predicted clusters match up with the clusters based on actual author characteristics. In other words, we use clustering to produce a summary of the types of authors in the data set and use the actual author characteristics as a baseline for evaluating the clusters produced using the predicted author characteristics.

This clustering is not a new task, on the one hand, because much work has already been done on profiling authors according to various biographical (Koppel *et al.*, 2003; Argamon *et al.*, 2009a,b; Mukherjee and Liu, 2010; Nguyen *et al.*, 2011; Sarawgi *et al.*, 2011) and political (Laver *et al.*, 2003; Yu *et al.*, 2008; Diermeier *et al.*, 2011; Dahllof, 2012; Iyyer *et al.*, 2014) attributes. On the other hand, the ability to cluster texts using predicted author profiles is a new approach not possible in previous work which (1) did not use as many profiles and (2) did not combine biographical and political profiles together. Thus, this clustering is an expansion of previous work both in terms of the number of profiles brought together and in terms of the clustering made possible by this larger number of profiles.

Clustering is done using the Simple Expectation Maximization, Simple K-Means and Cascade K-Means, and Canopy algorithms (Hall *et al.*, 2009). The similarity between clusters is measured using the Measure of Concordance (MoC; Pfitzner *et al.*, 2009). Given the newness of clustering texts according to multiple author profiles, we look at these four clustering algorithms in order to set up a baseline for further work in improving this sort of analysis.

The MoC ranges from 0 to 1, with 0 indicating no overlap between cluster membership and 1 indicating complete overlap. The measure is very sensitive to cluster differences and, as will be seen, tends to be quite low. The definition of the MoC measure is given in Table 8. It contains two components: first, the main measure of similarity, the overlap measure; second, a normalizing component to control for different numbers of clusters across the two clusterings. The number of instances shared between two clusters is represented by f_{ij}, where i and j represent specific clusters within a clustering. In other words, there are two clusterings, I and J. Each is made up of a number of individual clusters, for example i_1 and j_1. Thus, f_{ij} represents those instances which are present in both cluster 1 of set I and cluster 1 of set J.

The terms p_i and q_j represent the total number of instances contained in, respectively, cluster 1 of clustering I and cluster 1 of clustering J. The ratio f_{ij}/p_i thus represents the number of instances in cluster I_i shared with cluster J_j out of the total number of instances in cluster I_i. For each comparison between clusters, then, we have the probability that the instances are shared, in both directions: (f_{ij}^2/p_iq_j). This is because the probability of overlap given cluster I_i has been multiplied by the probability of overlap given cluster J_j. This symmetric probability of overlap between clusters is repeated for each combination of clusters within the two clusterings and then summed. The situation is complicated somewhat by the fact that two clusterings may contain different numbers of clusters. Thus, the measure is normalized by the geometric mean of the number of clusters in each clustering, as shown in Table 8. The result, then, is a measure of the similarity between two clusterings which performs well under various conditions (Pfitzner *et al.*, 2009) and is quite sensitive to differences between clusters. The algorithm for computing the MoC for any two clusterings is given in Table 9.

Table 8 Measure of concordance

$$(1/(\sqrt{IJ}-1))(\textstyle\sum_{\downarrow}(i=1)^{\uparrow}I=\sum_{\downarrow}(j=1)^{\uparrow}J=[(fij^{\downarrow}(2)\Big)/(p_{\downarrow}(i)q_{\downarrow}j)]-1)$$

Table 9 Algorithm for calculating the measure of concordance

1	Let I be the gold-standard clustering, with i_1, i_2... being each individual cluster
2	Let J be the text-based clustering, with j_1, j_2... being each individual cluster
3	For each possible combination of i_N and j_N :
4	Square the number of instances of i_N also contained in j_N
5	Multiply the total instances in $_iN$ and the total instances in j_N
6	Sum results from all possible combinations of clusters within each clustering
7	Normalize this sum with geometric mean of number of clusters within each clustering

Table 10 Similarity between clusterings using the MoC

	EM	K-Means	Cascade	Canopy
Individual	0.0161	0.0276	0.0443	0.0120
Aggregated	0.0283	0.0150	0.0561	0.0261

Clustering texts in a data set by biographical and political attributes of their author, as distinct from clustering the texts by content or topic or sentiment toward topics, gives us a birds-eye view of the individuals represented in a data set. In some cases where authorship analysis is applied (e.g. the study of literature) this capability is not important. However, in other cases (e.g. the study of large number of blogs), this capability is almost as important as analyzing individual authors because with such large numbers of authors that task becomes to analyze the crowd rather than the individuals. The test here is whether the clusterings produced using the predicted profiles (produced using textual features) are similar to the clusterings produced using the actual profiles (from the gold-standard meta-data, the ground-truth).

The similarity between gold-standard and predicted-class clusterings across algorithms is shown in Table 10, using the MoC as a measure of clustering similarity. The MoC is sensitive to clustering dissimilarity and the measures tend to be low. The Cascade K-Means algorithm performs best on both the individual and aggregated data sets. While these measures are somewhat low across the algorithms, this provides a first approach to evaluating this sort of clustering by author profiles.

4 Independence of Profiles

In this section we look closely at the predictive features for each of the seven well-performing profiles of author characteristics. First, we look at how correlated the feature weights are across profiles. Second, we look how topic-independent each profile is, by determining the ratio of topic-dependent to topic-independent predictive features.

4.1 Correlation of author profile feature weights

There are relationships between author characteristics in terms of the memberships of individuals, as we have seen above with the special interest group ratings. In other words, it is likely that if someone is a member of the Democratic party, they will fall on one side of the special interest group components. Another example of this is race: a non-white member of congress is more likely to be a member of the Democratic party than the Republican party in this data set. These are relationships between author characteristics themselves; a separate but related question is whether the features which can predict one characteristic are also used to predict a different characteristic.

In this section we look at the Pearson correlations between feature weights for each profile. The linear SVM produces weights for each classification, with the weights of one class located at 1 and the other at 1. Thus, weights closer to 1 indicate that a feature has more predictive power, with the sign indicating which class the feature predicts. Multi-class models (such as geography and religion) produce a set of feature weights for each pairwise combination of classes (e.g. predicting North versus South, North versus West). We are interested in the weight of each feature, then, not in its sign. We can think of the absolute values of the feature weights as the usefulness for that feature in making a given prediction. To simplify the discussion, we only consider the feature weights from the aggregated speeches data set.

Table 11 below shows the Pearson correlations for the absolute values of the feature weights for the well-performing social author characteristics. We see that all of the classifications are at least

Table 11 Pearson correlations between social profile feature weights, aggregated

	Age	Mil.	M-W	N-M	N-S	N-W	S-M	S-W
Age	–	0.350	0.253	0.234	0.207	0.224	0.233	0.248
Military	0.350	–	0.221	0.216	0.214	0.206	0.213	0.248
Midwest-West	0.253	0.221	–	0.372	0.221	0.409	0.403	0.442
North-Midwest	0.234	0.216	0.372	–	0.538	0.496	0.358	0.229
North-South	0.207	0.214	0.221	0.538	–	0.520	0.341	0.362
North-West	0.224	0.206	0.409	0.496	0.520	–	0.213	0.372
South-Midwest	0.233	0.213	0.403	0.358	0.341	0.213	–	0.347
South-West	0.248	0.248	0.442	0.229	0.362	0.372	0.347	–

Table 12 Pearson correlations between ideological profile feature weights, aggregated

	Party	Institutions	Animals	Chamber
Party	–	0.971	0.628	0.210
Institutions	0.971	–	0.650	0.216
Animals	0.628	0.650	–	0.246
Chamber	0.210	0.216	0.246	–

slightly related (between 0.2 and 0.3). Age and Military are more highly related at 0.350. This is in part related to the population of speakers, with older members of congress more likely to have served in the military. The correlations between Age and Geography range from 0.207 (North-South) to 0.253 (Midwest-West). This shows that these attributes are independent enough to include in the profile-based authorship method. The correlations between the different geography classifications are, as we would expect, sometimes high. Part of this is an artifact: any two classifications which share a single class (e.g. Midwest-West and South-Midwest) will be correlated because they share the same predictive features. Classifications which do not share a specific class remain low (e.g. Midwest-West and North-South at 0.221).

Correlations between feature weights for ideological profiles are shown in Table 12. Some of these are highly correlated, such as Party and the Government and Institutions special interest group rating (0.971). This correlation is high enough that there is no need to include both profiles for authorship analysis, given that they contain the same information. The other correlations, however, are low enough that each profile provides unique information about a text's author.

Table 13 High correlations between profiles

Profile 1	Profile 2	Aggregated	Individual
Party	Animals	0.628	0.676
Party	Institutions	0.971	0.945
Animals	Institutions	0.650	0.707
Animals	North-South	0.416	0.432

So far we have been concerned with the correlations between feature weights in order to determine which profiles should be used for authorship analysis, on the assumption that the profiles should be both well-performing and relatively independent. It is also insightful, however, to look at the correlations as a measure of similar patterns in language use across different social and ideological groups. If two profiles are not highly correlated, there is nothing interesting to report. However, it is important to ask why some profiles are highly correlated. To this end, Table 13 shows all of the high correlations between profiles (ignoring classifications with overlapping profiles). Only four pairings have correlations above 0.4. Three of these are expected (e.g. Party Membership and special interest group ratings). One which is not expected, however, is the high correlation (0.416 for aggregated speeches) between language patterns for authors who are for and against animal rights and who are from the North and the South.

4.2 Topic-independence of author profile feature weights

In this section we look at the degree of topic-independence of each profile by measuring the usefulness of both topic-dependent and topic-independent features. First, 'usefulness' is defined using the absolute values of the feature weights. Topic-independent features are defined as either (1) part-of-speech features or (2) function words (i.e. closed-class words). Topic-dependent features are defined as any feature which includes a lexical word (e.g. an open-class word). Table 14 shows the breakdown of features across these categories.

The single largest category is topic-dependent lexical words (3,037). While there are not many individual function words, there is a fair number (717) of function word features because of sequences of function words (e.g. 'will have been'). Over half of the features are topic-independent. This is not surprising given the evaluation during feature extraction which strongly favors features that are non-sparse across different topics.

The total topic-dependence of the social profiles is shown in Fig. 4. The absolute value of feature weights for each profile are summed across both topic-dependent and topic-independent features. This gives us a measure of the amount of information used for the classification that is contained by both sorts of features, which in turn tells us the relative topic-dependence of a given profile. The goal is to have topic-independent profiles which do not depend upon particular debates or contexts for their predictive powers. Measures are shown for both individual speeches and aggregated speeches. Each percentage represents that portion of predictive power contained by a given type of feature (e.g. 54.58% of the predictive power for the age characteristic comes from topic-dependent features). This use of feature weights is descriptive and not used to improve performance; however, future work could incorporate such feature information to improve the predictive power of the classifiers (Guyon *et al.*, 2002; Forman, 2003; Guyon and Elisseeff, 2003).

Table 14 Topic-independent and topic-dependent features by type

Type	Sub-type	Number
Topic-independent	Part-of-speech	2,873
Topic-independent	Function word	717
Topic-dependent	Containing lexical word	3,037

As shown in Fig. 4, topic-dependent features provide a majority of all the predictive power for each social profile. This is true even though there are more topic-independent features. Further, even though there is a wide range in total feature coefficients (e.g. weights, from 194.51 for Age to 409.15 for the South-Midwest classification), the relative distribution of predictive power is similar across the profiles. In Fig. 6 the percentage of the predictive power for each type of feature is shown for the individual speech data set and for the aggregated speech data set. We see that in every case the aggregated models

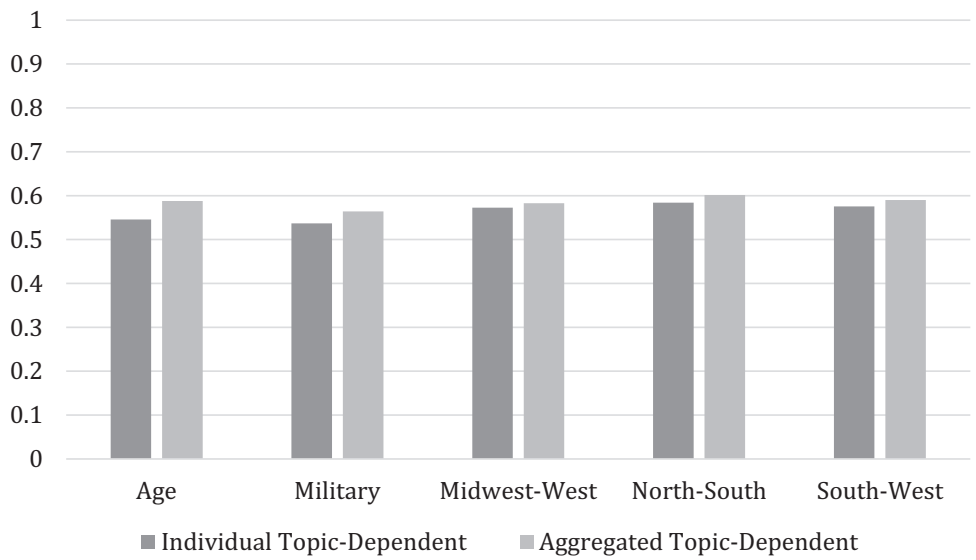

Fig. 4 Total topic-dependence of social profiles, individual/aggregated by percentage

have more predictive power in topic-dependent features than is the case for the individual speech models. This is because, as speeches are aggregated, topic-based features become less sparse and thus more useful for classification.

This measure of topic-dependence for the ideological profiles is shown in Fig. 5. Here the profiles are much more topic-dependent, ranging from 61.07 to 65.08%. And, again, the aggregated speech data set is more topic-dependent, although the difference between the two is smaller than above. This indicates that the ideological profiles, as we would expect, are more topic-dependent than traditional sociolinguistic characteristics like age. However, topic-independent features still have a significant predictive power for these profiles.

The point of this section has been to show that the predictive features for each of the profiles are (1) independent of one another and (2) relatively topic-independent. Both of these properties have been measured, showing the variations in both among profiles. This is an important task because it allows us to quantify the relationships between the profiles and the relationships between the predictions and particular topics.

A similar measure can be used to look at the potential predictive power of types of representations across the classes. Fig. 6 shows the relative predictive power for POS features (part-of-speech features; e.g., 'nn'), closed-class lexical features (e.g. function words), and open-class lexical features (e.g. non-function words). While the exact proportions vary across classes, the general hierarchy is clear: open-class items have the most predictive power, followed by part-of-speech representations with function words not far behind. This largely mirrors the topic-dependence information discussed above, with a further distinction between part-of-speech features and function words, both of which contribute significantly to the classification.

5 Profile-Based Authorship Analysis

Now that we have well-performing, unique, and relatively topic-independent profiles of the social and ideological characteristics of the authors we can use these profiles for authorship analysis of texts. Traditional authorship analysis consists of (1) attributing a given text to one of a set of known individuals or (2) determining whether two texts were written by the same individual (the fundamental

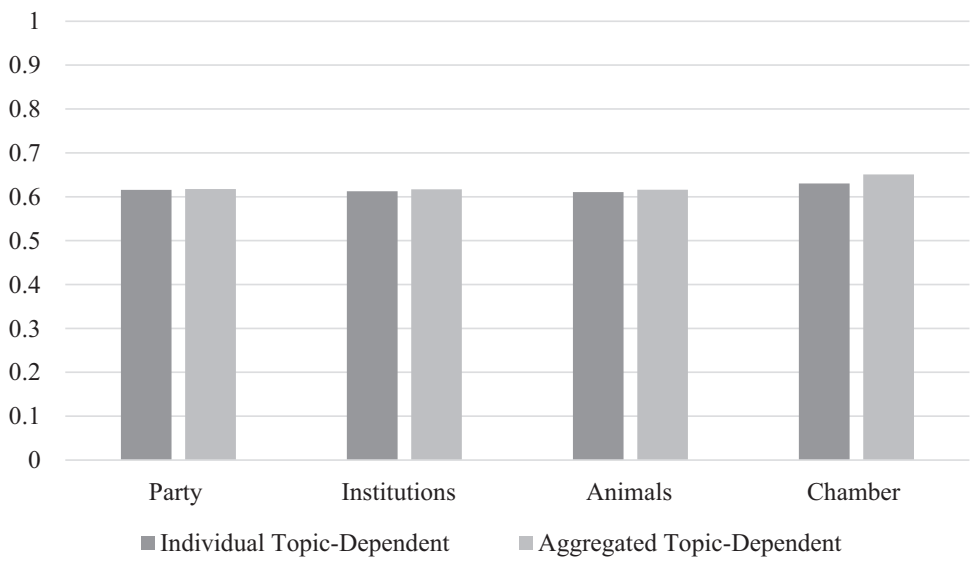

Fig. 5 Total topic-dependence of ideological profiles by percentage

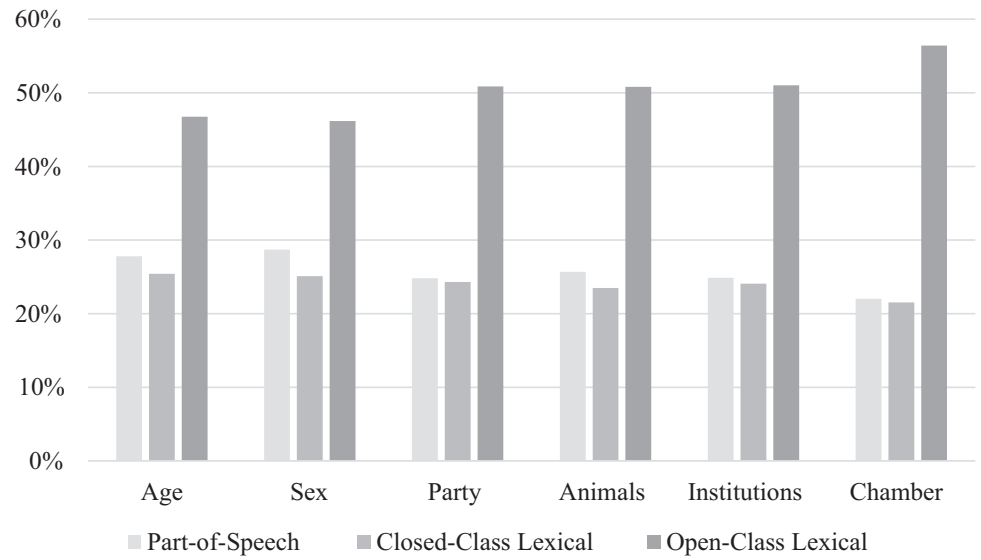

Fig. 6 Predictive power by feature type across classes by percentage

task of authorship analysis according to Koppel *et al.*, 2012). Profile-based authorship analysis allows us to expand these tasks. Here we undertake two additional tasks: (1) We use author profiles to build a text similarity measure which provides the probability that two texts were written by the same unknown individual and also allows us to see which profiles separate the two texts. (2) We use author profiles to cluster texts according to shared attributes of the authors; thus, rather than identifying individuals we use authorship analysis to identify texts by groups of similar individuals (as discussed in Section 3.1).

This profile-based method of authorship falsification is essentially an ensemble method (Opitz and Maclin, 1999; Rokach, 2010) which combines hierarchical classifier methods as parts of a larger ensemble-based classifier. In this case, each profile classification (e.g. sex) is an input to the ensemble system. The ensemble system, the profile-based falsification method described in this section, combines the results from these individual parts. This approach is also a data fusion approach, in that the ensemble is made up of classification by different sorts of latent authorial variables (e.g. age, sex, political party), rather than made up of applications of different classifiers to the same variables.

5.1 Profile-based author similarity measure

We measure author similarity by looking at how many of the individual profiles overlap. If the two speeches have different values for a particular profile, we know that they were written by different individuals. In other words, each profile without a shared value increases the dissimilarity between the authors of the two speeches. Of course, the profiles are not always accurate and, more importantly, vary in their accuracy. Thus, we weight each profile by its F-Measure. If a particular profile were always accurate, we would use a weight of 1. If, on the other hand, a particular profile has an F-Measure of 0.344, we would use a weight of 0.344. In this context a higher value means that the two speeches are more likely to have been written by different individuals. If a particular profile is always accurate and if two speeches differ on that profile, we know that the speeches were written by different individuals.

We sum these overlap scores for all of the profiles together to create a weighted measure of similarity such that lower values represent texts written by more similar individuals. This is a way of combining the information from multiple profiles which vary in accuracy in order to determine the similarity of the authors of two texts. The algorithm for

Table 15 Profile-based author similarity measure

1	Let T_1 be a text that is distinct from T_2
2	Let T_1 and T_2 have profile values $P1_i$, $P1_j$. . .. and $P2_i$, $P2_j$. . ..
3	For $P1_i$ and $P2_i$, $P1_j$ and $P2_j$. . .
4	$PD(P_i) = 1$ iff $P1_i \neq P2_i$
5	Let $F1(P_i)$ = F-Measure of classification by profile P_i
6	$wPD(P_i) = PD(P_i) * F1(P_i)$
7	Let ASD be the Author Similarity Measure
8	$ASD = \sum wPD(P_i), wPD(P_j)$. . ..

Table 16 Baseline implementation of Burrows' Delta

1	Let T_1. . .T_2. . . represent a set of distinct texts
2	Let F_1. . .F_2. . . represent a set of distinct text-based features
3	$RF(F_i)(T_j)$ = relative frequency of feature F_i in text T_j
4	Let N_t = number of texts in T_1. . .T_2. . .
5	$aRF(F_i) = \sum RF(F_i)(T_j)$. . . $RF(F_i)(T_k)$. . . / N_t [for F_i in all texts]
6	$Df(T_i)(F_j) = \lvert aRF(F_j) - RF(F_j)(T_i) \rvert$
7	Let $mDf(F_j)$ = mean value of $Df(F_j)$ in all texts
8	Let $stdDf(F_j)$ = standard deviation of $Df(F_j)$ in all texts
9	Let $StandardDf(T_i)(F_j) = [Df(T_i)(F_j) - mDf(F_j)]$ / $stdDf(F_j)$
10	$\Delta T_i = \sum StandardDf(T_i)(F_j)$. . . $StandardDf(T_i)(F_k)$. . .. [for each F in T_i]

computing this measure is shown in Table 15. It is possible, of course, that two different individuals will have all of the same characteristics. For example, two liberal white male senators from the north will have all the same speaker characteristics, and thus will not be identified as different individuals even though they are. Therefore, this measure can tell us if two speeches have different authors or if two speeches were produced by similar individuals (not necessarily the same individual). The use of multiple profiles is similar to the unmasking method of authorship verification (Koppel *et al.*, 2007), in that the more profiles two texts differ by the deeper their dissimilarity and the more likely they were produced by different individuals.

The Author Similarity Measure was tested on a data set consisting of pairs of profiles of speeches. The data set was formed by taking each profile of each speech in the test set and combining it with (1) a different speech by a different individual and (2) a different speech by the same individual. Thus, for the data set of individual texts, there are 31,289 pairs, half of which share the same author. For the aggregated data set, because there is only one document per author in the test set, the test set was combined with predicted profiles from the 108th congress, creating a data set of 946 pairs of speeches, half of which share the same author. In this way the task does not compare texts against a range of texts by candidate authors but rather represents a single comparison to determine if the texts share a single author, an important difference (Schaalje *et al.*, 2011).

We use a version of Burrows' Delta (Burrows, 2002) as a baseline for the Author Similarity Measure. This Delta is calculated using the same feature set discussed in Section 2.1. The pseudo-code is shown in Table 16. The main difference between this and other versions of Burrows' Delta is the feature set upon which it is computed. As discussed in Argamon (2008), this measure really places texts on a scale of difference, which means that we can set thresholds for rejecting shared authorship, as in the profile-based measure described above. Both measures differ from pure text similarity measures (Forsyth and Sharoff, 2013) in that the standard for classification is known author characteristics rather than annotated properties of texts (e.g. similarity of topic or structure). Thus, Burrow's Delta was chosen as a baseline, instead of text similarity measures such as cosine distance.

We have seen above that some profiles are much more accurate than others, in part because of the skewed nature of some author characteristics in this data set. Here we use all profiles: Age, Chamber, Geography, Previous Military Service, Party Membership, Race, Religion, Sex, (SIG refers to components of special interest group ratings). The idea is that even if a particular profile is not entirely accurate we can nonetheless make use of the new information that it does provide by appropriately weighting how important that information is (here, operationalized by weighting by the profile's F-Measure). Thus, we want to leverage profiles of all performance levels for the task for authorship analysis.

Tables 17 through 20 show the results of authorship identification with each of the measures of author similarity, one direct (Burrows' Delta) and

one mediated by author profiles (Author Similarity Measure). The threshold indicates the level of difference above which two texts are considered to be written by different authors. For Burrows' Delta, the Individual and Aggregated data set results are reported separately because the required thresholds do not overlap. The data set, as mentioned above, includes either 31,289 (individual speeches) or 946 (aggregated speeches) pairs of texts, both with a total of 535 speakers. Further, these speakers are similar in many respects, requiring fine-grained distinctions to be made in order to differentiate texts.

First, we see in Table 17 that aggregated speeches perform better than individual speeches for this task, as we would expect from the profile-building results. The aggregated data set achieves an F-Measure of 0.644 (threshold at 2) and the individual data set achieves an F-Measure of 0.592 (again, threshold at 2). The highest aggregated result is with all profiles used while the highest individual result (F-Measure 0.611) is with only the select profiles used, as shown in Table 18. This is likely because, while somewhat poor performing, all of the profiles on the aggregated data set contribute some amount of information, however small, while that is not the case on the individual data set where some profiles perform poorly enough to add no new information. The thresholds are lower on the data set with only the select profiles used, given that the there are fewer profile overlaps contributing to the measure.

Table 19 shows the performance of Burrows' Delta, calculated as described in Table 16, on the task using the individual data set. The highest F-Measure is 0.515 (threshold at 0.05, marking small differences between texts). Table 20 shows the performance of Burrows' Delta on the aggregated data set, with a highest F-Measure of 0.633 (threshold at 0.25).

Thus, the profile-based method out-performs the Delta measure on this large data set in both cases. On aggregated texts, for which the features are not sparse, the performance is close: 0.644 versus 0.633 F-Measure. In this case, then, without sparse features, the profile-based method is better but not by a large margin. On the individual speech data set, however, which has very sparse features by necessity, the profile-based method outperforms Burrows' Delta by a wide margin, 0.611 versus 0.515 F-Measure. To put this in perspective, the baseline F-Measure for random guesses is 0.500, showing that Burrows' Delta does not improve performance much above the baseline in this situation. Thus, we see that the profile-based method performs well above baseline on a data set with a large number of speakers, a large number of texts, and a large number of features, even when profiles vary widely in their individual performance.

Table 17 Performance of author similarity measure with all profiles, individual/aggregated

Threshold	Precision	Recall	F-measure
1	0.616/0.659	0.563/0.631	0.508/0.613
2	0.595/0.645	0.593/0.644	0.592/0.644
3	0.599/0.650	0.576/0.610	0.550/0.582

Table 18 Performance of author similarity measure with select profiles, individual/aggregated

Threshold	Precision	Recall	F-measure
0.75	0.617/0.586	0.607/0.578	0.599/0.567
1	0.611/0.645	0.611/0.626	0.611/0.614
2	0.626/0.648	0.549/0.541	0.467/0.440

Table 19 Performance of Burrows' Delta on same task, individual

Threshold	Precision	Recall	F-measure
0.20	0.536	0.495	0.342
0.15	0.520	0.498	0.381
0.10	0.523	0.510	0.463
0.05	0.514	0.515	0.515

Table 20 Performance of Burrows' Delta on same task, aggregated

Threshold	Precision	Recall	F-measure
0.25	0.665	0.640	0.633
0.50	0.646	0.567	0.519
0.75	0.643	0.535	0.455
1.00	0.640	0.511	0.401

5.2 Evaluation on held-out test set

Now that we have optimized the SVM parameter settings, investigated feature relatedness between profiles and topic-dependence of profiles, and optimized settings for authorship falsification, we turn to a new test set in order to apply the entire system to unseen data. This new test set consists of speeches from the 110th to 112th congresses (2007–13), with the descriptive statistics given in Table 21. Only speeches given by speakers present in earlier congresses are used. The main reason for using an additional test set later in time than the training data (in this case, separated by the training data by at least 2 years) is to see how well the performance of the optimized system holds up in new time periods and new political contexts. This is used instead of cross-validation for two reasons: first, the predicted profiles must be used for authorship falsification, so that multiple divisions between testing and training data is not possible (e.g. predictions from only one-fold could be used for falsification; but it is not clear which fold should be used); second, this allows us to simulate how the system will perform over time when applied to new data which does not yet exist (e.g. the 113th congress).

Table 21 Descriptive statistics for held-out test set

Speeches	Period	Instances	Average words	Range of words
Aggregated	Test: 110–112	1,573	26,236	500–528,531
Individual	Test: 110–112	27,538	1,148	500–3,729

The models built using the 104th through 108th congresses were applied to the 110th through 112th congresses, with the resulting F-Measures shown in Fig. 7. The same features were used (e.g. no new feature evaluation). This represents, then, the performance of the original models on data even further separated in time. Overall, the same patterns continue to hold: predictions about aggregated speeches outperform predictions about individual speeches; political attributes outperform biographical or social attributes. Overall, this shows consistent performance across time.

These predicted profiles are then used for the task of authorship falsification with Burrow's Delta as a baseline. Both approaches are performed using the optimal settings as evaluated above, with the results shown in Fig. 8. As before, the test set is formed by creating evenly divided sets of pairs of speeches: half are pairs of speeches which share the same author (but are not the same speech), and half are pairs of speeches which do not share the

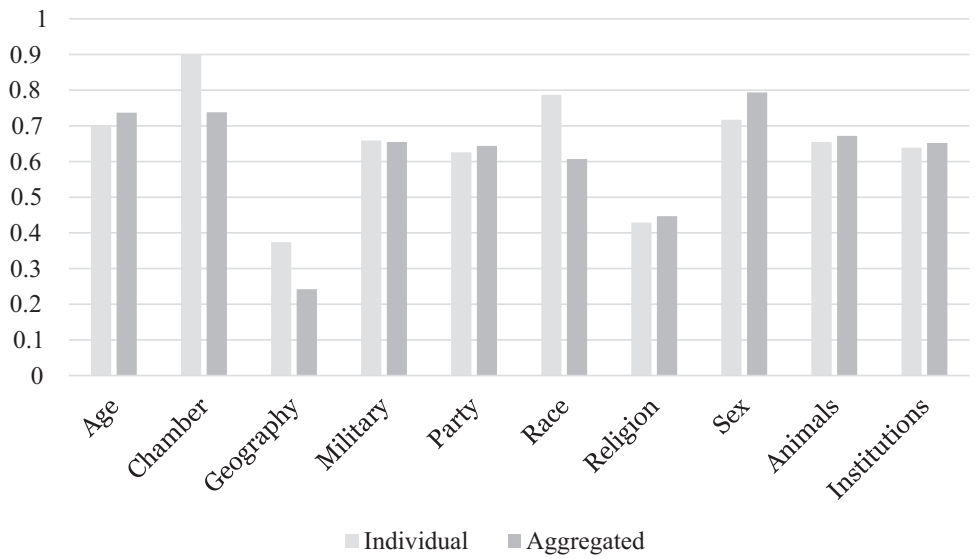

Fig. 7 Performance by F-measure, 110th–112th congresses

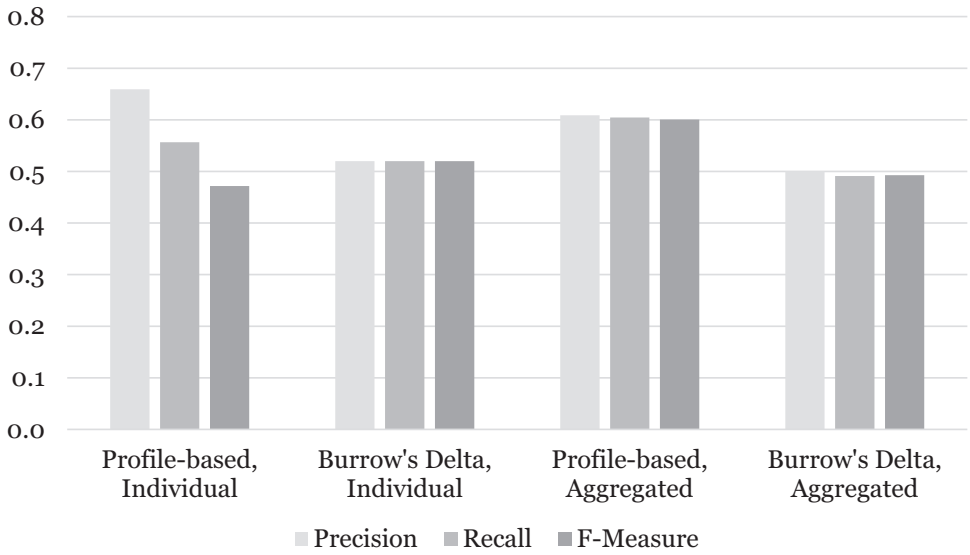

Fig. 8 Authorship falsification on 110th to 112th congresses

same author. This creates a test set of 54,700 pairs of speeches for the individual data set and 2,800 pairs of speeches for the aggregated data set. First, Burrow's Delta does not perform above the majority baseline of 0.5 F-Measure on either data set. The profile-based approach is above baseline slightly for recall and more so for precision but below baseline for F-Measure on the individual data set. As before, then, the profile-based approach is just slightly better than Burrow's Delta at the level of individual speeches. On aggregated speeches, however, the profile-based approach outperforms Burrow's Delta by a wide margin. Again, although the performance is slightly lower than before, this shows overall a consistent performance of the approach even on new data (with profiles built with models from the 104th through 108th congresses).

6 Conclusions

The ensemble profile-based authorship analysis method described in this article out-performs existing methods for the task of determining shared authorship of texts and supports the clustering of texts by author characteristics. In addition, this article has shown that a range of n-gram windows and feature types can be combined without creating overly sparse vectors and that both the mutual dependence and the relative topic-dependence of profiles can be measured, an important prerequisite for profile-based methods. This represents a step toward untangling the large number of factors which combine to produce the linguistic variations which allow authorship analysis in the first place.

Acknowledgements

This project was supported by a grant from the Intelligence Community Postdoctoral Research Fellowship Program. All statements of fact, opinion, or analysis expressed are those of the author and do not reflect the official positions or views of the Intelligence Community or any other U.S. Government agency. Nothing in the contents should be construed as asserting or implying U.S. Government authentication of information or Intelligence Community endorsement of the author's views.

References

Abbasi, A. and Chen, H. (2008). Writeprints. *ACM Transactions on Information Systems*, **26**(2): 1–29.

Antonia, A., Craig, H., and Elliott, J. (2013). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution. *Literary and Linguistic Computing*, **29**(2): 147–63.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, **23**(2): 131–147.

Argamon, S., Cooney, C., Horton, R., Olsen, M., Stein, S., and Voyer, R. (2009a). Gender, race, and nationality in black drama, 1950-2006: mining differences in language use in authors and their characters. *Digital Humanities Quarterly*, **3**(2).

Argamon, S., Goulain, J., Horton, R., and Olsen, M. (2009b). 'Vive la différence!' Text mining gender difference in French literature. *Digital Humanities Quarterly*, **3**(2).

Burrows, J. (2002). Delta: A measure of stylistic difference and guide to likely authorship. *Literary and Linguistic Computing*, **17**(3): 267–87.

Dahllof, M. (2012). Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches: a comparative study of classifiability. *Literary and Linguistic Computing*, **27**(2): 139–53.

Diermeier, D., Godbout, J. -F., Yu, B., and Kaufmann, S. (2011). Language and ideology in Congress. *British Journal of Political Science*, **42**(1): 31–55.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, **3**: 1289–305.

Forsyth, R. S. and Sharoff, S. (2013). Document dissimilarity within and across languages: a benchmarking study. *Literary and Linguistic Computing*, **29**(1): 6–22.

Garera, N. and Yarowsky, D. (2009). *Modeling Latent Biographic Attributes in Conversational Genres. Proceedings of Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, 710–18.

Gentzkow, M. and Shapiro, J. (2013). *Congressional Record for 104th-109th Congresses: Text and Phrase Counts*. Ann Arbor, MI, ICPSR33501-v2.

Grieve, J. (2007). Quantitative Authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, **22**(3): 251–70.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**: 1157–82.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**: 389–422.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, **11**(1): 10.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, **22**(4): 405–17.

Hoover, D. (2004). Testing Burrow's Delta. *Literary and Linguistic Computing*, **19**(4): 453–75.

Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. (2014). Political Ideology Detection Using Recursive Neural Networks. Proceedings of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, Association for Computational Linguistics. pp. 1113–22.

Jockers, M. L. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, **25**(2): 215–23.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, **13**(3): 637–49.

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

Koppel, M., Argamon, S., and Shimoni, A. (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, **17**(4): 401–12.

Koppel, M., Schler, J., Argamon, S., and Winter, Y. (2012). The 'fundamental problem' of authorship attribution. *English Studies*, **93**(3): 37–41.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research*, **8**: 1261–76.

Laver, M., Benoit, K., Garry, J., and Trinity, K. B. (2003). Extracting policy positions from political texts using words as data. *The American Political Science Review*, **97**(2): 311–31.

Luyckx, K. and Daelemans, W. (2010). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, **26**(1): 35–55.

Mukherjee, A. and Liu, B. (2010). *Improving Gender Classification of Blog Authors. Proceedings of the Conference on Empirical Methods in Natural Language*

Processing. Stroudsburg, PA: Association for Computational Linguistics. pp. 207–17.

Nguyen, D., Smith, N. A., and Ros, C. P. (2011). *Author Age Prediction from Text using Linear Regression. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Stroudsburg, PA: Association for Computational Linguistics. pp. 115–23.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, **11**: 168-98.

Pfitzner, D., Leibbrandt, R. and Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, **19**(3): 361–94.

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In Schoelkopf B., Burges C., and Smola A. (eds), *Advances in Kernel Methods - Support Vector Learning*. Cambridge: MIT Press. pp. 185–208.

Poole, K. and Rosenthal, H. (2007). *Ideology and Congress*. Edison, NJ: Transaction Publishers.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, **33**: 1–39.

Sarawgi, R., Gajulapalli, K., and Choi, Y. (2011). *Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Stroudsburg, PA: Association for Computational Linguistics. pp. 78–86.

Schaalje, G. B., Fields, P. J., Roper, M., and Snow, G. L. (2011). Extended nearest shrunken centroid classification: a new method for open-set authorship attribution of texts of varying sizes. *Literary and Linguistic Computing*, **26**(1): 71–88.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, **3**(5): 38–56.

Thomas, M., Pang, B., and Lee, L. (2006). *Get out the Vote: Determining Support or Opposition From Congressional Floor-Debate Transcripts. Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics. pp. 327–35.

Yu, B. (2013). Language and gender in Congressional speech. *Literary and Linguistic Computing*, **29**(1): 118–32.

Yu, B., Kaufmann, S., and Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology and Politics*, **5**(1): 33–48.