Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability

Mats Dahllöf
Department of Linguistics and Philology, Uppsala University, Sweden

Abstract

The present study explores automatic classification of Swedish politicians and their speeches into classes based on personal traits—gender, age, and political affiliation—as a means for measuring and analyzing how these traits influence language use. Support Vector Machines classified 200-word passages, represented by binary bag-of-word-forms vectors. Different feature selections were tried. The performance of the classifiers was assessed using test data from authors unseen in the training data. Author-level predictions derived from twenty-one text-level predictions reached an accuracy rate of 81.2% for gender, 89.4% for political affiliation, and 78.9% for age. Classification concerning each basic distinction was applied to general populations of politicians and to cohorts defined by the other classes. The outcomes suggest that the extent to which these personal traits are expressed in language use varies considerably among the different cohorts and that different traits affect different layers of the vocabulary. The accuracy rates for gender classification were higher for the right wing and older cohorts than for the opposite ones. Age prediction gave higher accuracy for the right wing cohort. Political classification gave the highest accuracy rates when all forms were included in the feature sets, whereas feature sets restricted to verbs or function words gave the highest scores for gender prediction, and the lowest ones for political classification.

Correspondence:
Mats Dahllöf, Department of Linguistic and Philology, Uppsala University, Box 635, 751 26 Uppsala, Sweden.
E-mail:
mats.dahllof@lingfil.uu.se

1 Introduction and Purpose

The present study explores author-level and text-level automatic classification of politicians and their speeches into classes based on personal traits as a means for measuring and analyzing how these traits influence language use. This method will be applied to three different binary personal trait distinctions: gender (or sex, FM), i.e. female (F) versus male (M), political affiliation (LR), i.e. left-wing (L) versus right-wing (R), and age (OY), i.e. older (O) individuals (birth $\leq$ 1953) versus younger (Y) ones (birth $\geq$ 1959). (The letters F, M, L, R, O, and Y will be used to denote the six classes.) The study draws its data from speeches delivered in the Swedish parliament during the seven

Literary and Linguistic Computing, Vol. 27, No. 2, 2012.

doi:10.1093/llc/fqs010 Advance Access published on 6 April 2012

annual sessions 2003/2004–2009/2010. The methodological idea is, to be more specific, that the classifiability of texts and authors into different classes reflects how much impact membership in these classes has on the text features employed for training and classification.

Classification into the three basic distinctions for general populations will be addressed, as well as for cohorts restricted to politicians belonging to one of the four other classes, e.g. gender classification as applied to the cohort of left-wing politicians. The abbreviation FM[L] will be used for this task. This gives us fifteen binary classification tasks, three unrestricted ones and twelve restricted ones, as shown in Table 1. If we construct classifiers for the fifteen tasks in a uniform way, their performance scores will reflect how difficult the various classification tasks are, relatively to the classifier design principle. This allows us to address two related sets of questions about politicians' use of language in the Swedish parliament:

- How strong is the classifiability of each basic distinction in the restricted cases compared to each other and to the unrestricted case? For instance, is gender (FM) easier to predict for the right-wing cohort (FM[R]) than for the left-wing group (FM[L])? (We will end up with a positive answer.)
- How well will classifiers predicting membership in the basic classes perform based on the presence or absence of different selections of word forms in short (200-word) stretches of text? This will allow us to see to what extent different personal traits affect different layers of the vocabulary.

These questions will be addressed by means of experiments with conventional state-of-the-art text classifiers. The focus will not be on sophisticated classifier engineering.

2 Background and Previous Studies

Sociolinguistic research has paid considerable attention to the ways in which personal traits are reflected in language use. They are also interesting from the point of view of automatic text classification in computational linguistics as examples of categories which are non-trivial to predict both for humans and machines.

Prediction of personal traits in authors based on spoken and written document data is a field which offers a wealth of possibilities. There are many kinds of characteristics that can be studied, e.g. gender, age, personality traits such as extraversion and neuroticism (Kim and Daelemans, 2008), political affiliation, and ideological orientation. Moreover, studies in this field may focus on any genre of linguistic performance. A key parameter is the length of the basic documents, which, as we will see below, has fallen in the range of single sentences to book-length texts. The formal design of the overall system, which typically involves a text representation scheme and basic classifier modules, also offers a large number of options. Documents are typically represented by high-dimensional vectors, defined through a process of feature construction, selection, and weighting. Furthermore, there are many machine learning algorithms for classifier generation, each of which can be modified by different parameter settings.

As can be expected, most studies in text classification targeting personal traits have focused on gender, whose status as a factor in human affairs

Table 1 The fifteen classification tasks

	Restricted to					
Unrestricted	**F(emale)**	**M(ale)**	**L(eft-wing)**	**R(right-wing)**	**O(lder)**	**Y(ounger)**
FM (female/male)	–	–	FM[L]	FM[R]	FM[O]	FM[Y]
LR (left-wing/right-wing)	LR[F]	LR[M]	–	–	LR[O]	LR[Y]
OY (older/younger)	OY[F]	OY[M]	OY[L]	OY[R]	–	–

is a central issue in many academic fields. The relation between gender and language has given rise to a wide-reaching discussion in general linguistics, but the present overview will only be concerned with classification-oriented research.

Koppel *et al.* (2003), in an early study on fiction and non-fiction prose, trained a Balanced Winnow classifier to tell texts by female authors from those written by men. They drew their data, 566 formal written documents, with a mean length of 34,000 words, from a genre-controlled and gender-balanced subset of the British National Corpus. Using few and content-neutral features weighted by relative frequency, they predicted author's gender with about 80% accuracy rate, a score based on fifty-six-fold cross-validation. Argamon *et al.* (2009) performed gender classification on historical book-length (mean length 75,000 words) French texts from the 12th to 20th Centuries. Using a Support Vector Machine (SVM) classifier (see Section 4), they found that surface word form features, which gave accuracy rates around 90%, were better than features based on lemmas or part-of-speech tags.

Boulis and Ostendorf (2005) studied gender classification for transcribed telephone conversations, whose mean length was 1,700 words for two speakers. Using SVM models and 120,000 bigram features with binary weighting to represent documents, they reached an accuracy rate of 92.5% when 14,969 conversation sides—equal numbers for each gender—were used for training. They also report that the gender of the addressee is important: In male–male and female–female conversations the speaker's gender is easier to classify, whereas they found evidence of some 'convergence of male and female linguistic patterns in cross-gender conversations'.

Studies on gender classification have often drawn their data from computer-mediated communication genres, e.g. e-mails, chat messages, weblogs, and online forum posts. Corney *et al.* (2002) showed that SVM models based on content-neutral features are useful for performing gender classification on e-mails in English. Kucukyilmaz *et al.* (2006) reported 84% accuracy in predicting author's gender for chat messages in Turkish. A Naive Bayesian classifier, in conjunction with a variety of content-neutral features, gave the highest accuracy rates. The classifier was trained on data from 100 female and 100 male individuals. Each 'document' was a selection of 3,000 words.

Gender classification on weblogs in English was studied by Schler *et al.* (2006). Their data, which were balanced gender-wise, comprised around 37,000 weblogs, each one on average comprising around 8,000 words. They evaluated Multi-Class Real Winnow classifiers by tenfold cross-validation: 502 stylistic features gave around 77% accuracy; 1,000 unigrams, those with highest information gain (see Section 4), intended to capture content, gave around 73% accuracy. The two sets combined reached 80.1%.

Otterbacher (2010) compared the values of style, content, and other features for gender classification of online movie reviews (mean length 256 words). She derived twenty content features by means of Latent Semantic Analysis. Employing the 15,650 reviews in the corpus for tenfold cross-validation, she found that style and content feature sets each give about 65% accuracy, using a logistic regression classifier.

Opsomer *et al.* (2008) studied gender classification in transcriptions of 1.9–3.0 year old children's speech (mean document length 616 words) reaching the best results, an accuracy of 70.5% (girl documents being slightly more numerous). They used an SVM bag-of-words approach, training classifiers on 805 documents. Gender classification has even been applied to fictional characters (Hota *et al.*, 2006). Komiya *et al.* (2009) found that a small set of features makes it possible to predict the suitable speaker's gender for single sentences in Japanese. Their classifier was almost as good as people in deciding the speaker's gender for sentences drawn from novels.

These classification studies indicate that both content-specific and more content-neutral features provide useful information in gender prediction. The author's gender is reflected in linguistic features regardless of which kind of discourse has been studied. However, it seems that gender classification is more difficult than topic classification, unless we look at texts where gender idiosyncrasies are more freely expressed, such as telephone conversations.

Several studies have found that words denoting family relations are strongly associated with female speech (Schler *et al.*, 2006; Nowson, 2006; Argamon *et al.*, 2007; Zhang *et al.*, 2009; Otterbacher, 2010), especially words relating to the concept of *child* (Boulis and Ostendorf, 2005; Sabin *et al.*, 2008; Argamon *et al.*, 2009). However, some family words are reported to be associated with male discourse (Sabin *et al.*, 2008). Among word forms which has been reported as typical of male discourse we find, for instance, swear words (Boulis and Ostendorf, 2005), words relating to politics, games, the Internet (Argamon *et al.*, 2007), money, sports (Schler *et al.*, 2006), and religious terms (Zhang *et al.*, 2009).

Political orientation prediction has been the subject of a number of text classification studies. The focus is typically on text classification, rather than on the political orientation as a trait of the author. Diermeier *et al.* (2012) (also see Yu *et al.*, 2008) used text classification to study legislative discourse. They used data from the US Senate treating each senator's complete set of speeches given during one Congress (2 years) as one document. (This is in sharp contrast with the present study, which is concerned with 200-word stretches of text.) Training data were drawn from the 101st to 107th Senates and the test data from the 108th Senate. They restricted their study to the twenty-five most conservative and twenty-five most liberal senators in each senate. This means that most of the senators from whom test data were drawn also provided training data. Their classifiers predicted political orientation—extremely conservative or extremely liberal—with 92% accuracy. Other contributions to political text classification are Greevy and Smeaton (2004), who classified websites into racist, neutral, and anti-racist ones, and Koppel *et al.* (2009), who predicted the affiliation of websites to specific Islamic organizations and ideologies. Jiang and Argamon (2008) report having improved accuracy in classifying weblogs as liberal or conservative by using features based on opinion mining.

Classification of age in authors has attracted less attention than gender and political orientation. Opsomer *et al.* (2008) report an accuracy rate of 80.5% telling children aged 1.9–2.4 years from those aged 2.4–3.0 years. The models of Schler *et al.* (2006) (described above), when applied to age classification into three age classes 13–17, 23–27, and 33–42 years, gave performance scores comparable to those for gender classification.

A methodological problem with a couple of studies exploring gender classification may be relevant to point out, since they reported remarkable accuracy rates. Both Nowson and Oberlander (2006) (also see Nowson, 2006), who studied personal weblogs in English, and Zhang *et al.* (2011), who drew their data from religious web forum posts, achieved accuracy rates above 90%. Furthermore, their systems were based on remarkably small feature sets (8 and 640 features, respectively). These had been selected out of large sets of candidate features, *n*-grams, and unigrams and bigrams, respectively. However, as Nowson and Oberlander (2006) admit, there is 'room for overfitting since the feature set is used to classify the very data from which it was derived'. There is reason to suspect that this is indeed the explanation for the high accuracy rates of their classifiers, as the average length of the documents (sequences of posts) was several thousand words, and the number of individuals in their data sets small, 71–100 authors. Mukherjee and Liu (2010) seem to adopt a similar method, 'mining'—from the full data set—part-of-speech sequences, which they assume 'represent true regularities'. The three studies seem to suffer from the same flaw: if the feature sets have been selected in a such a way that they are tailored to the full collection of data, prior to the cross-validation, the feature engineering itself or the overall classification method is not at all evaluated on unseen data. If the author set is small and the set of candidate features large, it is quite likely that it is possible to find a small subset of features that coincidentally does the trick for just that data set.

3 Data

The speeches held in the Swedish parliament during the seven annual sessions 2003/2004–2009/2010 were downloaded from the Swedish parliament website, http://www.riksdagen.se, where they

appear as published in the Hansard. The total collection amounts to about 30 million words and represents almost 600 speakers. The transcriptions are intended to be verbatim to the speeches as actually worded in the chamber, but are professionally edited and standardize morphology and syntax. The speeches, delivered by ministers and by ordinary members of parliament, represent different kinds of debate, e.g. deliberations leading to a decision and more general debates. Since the speakers often read from manuscripts, large portions of the corpus are prose-like. There are also passages that have the characteristics of spontaneous speech. However, the setting is highly formal and the communicative style monologic rather than interactive.

In order for the data to be as uniform as possible, the basic text documents for the present study are excerpts comprising the initial 200 words of the speeches. This decision was part of a series of choices regarding the selection of corpus data, as it will allow us to find, for each classification task, at least seventy-seven politicians for each class who individually delivered at least twenty-one documents and together at least 6,000 documents (see Section 4.2). At the same time, the excerpts are long enough to provide substantial information about the speaker's treatment of the subject matter, while being rhetorically alike as they are located as introductory portions. They almost always begin with the formal default address *Herr/Fru Talman*! ('Mr./Madam Speaker!'). The introduction also often makes explicit who is the 'real' addressee. Between 2 and 12 min are allotted for speeches. The shortest ones have the status of a reply to a previous speech, but are often longer than 200 words and will consequently be included in the data sets. The corpus contains 63,023 speeches of the required length. The number delivered by each politician varies considerably, between just 1 and 925.

Gender and age data were retrieved from member presentations at the parliament website. The party label appears together with the speaker's name for each speech in the Hansard. The distinction between left-wing and right-wing political affiliation follows the strongly established cooperative patterns of Swedish national politics. Left-wing affiliation is defined as membership in the parties Socialdemokraterna (s), 'The Social Democratic Party', Miljöpartiet (mp), 'The Green Party', and Vänsterpartiet (v), 'The Left Party'. The right-wing parties are Centerpartiet (c), 'The Centre Party', Folkpartiet liberalerna (fp) 'The Liberal Party', Kristdemokraterna (kd), 'The Christian Democrats', and Moderata samlingspartiet (m), 'The Moderate Party'. (s) was in government, supported by (mp) and (v), 2003/2004–2005/2006. The country was governed by a coalition which included (c), (fp), (kd) and (m) 2006/2007–2009/2010. As regards age, a distinction was made between older (O) politicians, born in 1953 or earlier, and younger (Y) ones, born in 1959 or later, excluding those born in 1954–1958. This choice was motivated by a wish to keep the two classes roughly equal in size over the different cohorts. The authors' actual age when delivering a speech included in the data set, was at least forty-nine, for the O group, and at most fifty-one years for the Y group.

4 Method

The setup for the present study involves two kinds of classification. First, there is the more basic issue of training and applying SVM models for text classification. A second stage is concerned with deriving author-level predictions based on the output from a text-level classifier for several texts. The selection of training data is a two-step procedure. First, a balanced set of training authors is selected. After that, a balanced training set of documents by these authors is compiled.

4.1 Text-level classifier design

The basic text classifiers used here are based on a straightforward application of state-of-the-art tools, foremost among which are SVMs. Classifiers belonging to this family were a natural choice for many of the studies described in Section 2 (Corney *et al.*, 2002; Nowson and Oberlander, 2006; Kobayashi *et al.*, 2007; Argamon *et al.*, 2009; Zhang *et al.*, 2009). Several studies (Boulis and Ostendorf, 2005; Opsomer *et al.*, 2008; Sabin *et al.*, 2008) compared SVM models with other

Table 2 The five feature selection schemes

all5k	Using the IG-topmost 5,000 features, with no other filtering, i.e. admitting proper nouns.
lx5k	Using the IG-topmost 5,000 features removing proper nouns.
lx2k	Like *lx5k*, but only the IG-topmost 2,000 features.
vb2k	The IG-topmost 2,000 verb forms. This scheme admits only those graph words which in the majority of cases and in at least three instances occurs with the part-of-speech tag VB (verb) in the Swedish Stockholm Umeå corpus (SUC) (Ejerhed *et al.*, 2006).
fu1k	Using only function words whose frequency is different in the two classes. (Given the setup, this implies a difference also in *relative* frequency.) The function word dictionary was also SUC based, counting as function words those graph words which in the SUC corpus carry either of the AB, DT, HA, HD, HP, IE, KN, PN, RG, RO, or SN part-of-speech tags in the majority of their instances. (This scheme gives less than 1,000 features.)

classifiers and concluded that SVMs are the best ones for gender prediction.

The present study relies on the SVM implementation of Joachims (1999), the SVMlight package (http://svmlight.joachims.org), version 6.02 (2008). A linear kernel and a biased hyperplane (the default setup) were used, as other options seemed to give lower accuracy rates. Training and application of the SVM models concern 200-word segments of speeches. Each such document is modeled by a vector of features, whose weighting is binary (or boolean), i.e. weight one expresses presence of a feature and zero its absence. A vector is consequently equivalent to a set of features. These are based on different selections of graph words, neutralizing letter case. Each data item is a vector labeled as a positive example, i.e. as belonging to the target class, or as a negative one. A sequence of labeled vectors provides training data for the SVM learning module, which produces the actual SVM. The SVM, when applied to a vector, returns a real number. This is the value of the decision function, whose sign—positive or negative—represents the classification verdict of the SVM. The absolute value represents the distance of the vector from the discriminating hyperplane. The author-level classification—see Section 4.3—is based on the assumption that this value reflects the probability of the prediction (cf. Pion and Hamel, 2007).

The classification experiments exploit a number of feature selection schemes, designed to shed light on which layers of the vocabulary contain the most useful features for the various classification tasks. They are described in Table 2. Word forms were case-neutralized, but no linguistic analysis was performed.

The selection schemes make use of a simple information-theoretic metric, *information gain* (IG), which is commonly used for finding the most significant features (Joachims, 1998; Boulis and Ostendorf, 2005; Schler *et al.*, 2006; Zhang *et al.*, 2009). In binary classification, IG for a binary feature and a class can be computed by this formula (Yang and Pedersen, 1997):

$$\mathrm{IG(F, C)} = \sum_{f \in \{F, \overline{F}\}} \sum_{c \in \{C, \overline{C}\}} P(f, c) \log_2 \frac{P(f, c)}{P(f)P(c)}$$

where F is the event that the feature F is present and C the event that the class C is exemplified. IG quantifies the information-theoretic value in bits of knowing whether a feature is present or not, given a certain classification task. A potential weakness of using this metric for feature selection is that is does not say anything about the joint value of a set of features, which, for instance, might be more or less independent. As one bit equals a decision in a binary classification task, the features we encounter in this study are worth considerably less than that. (Some examples are given in Table 6.)

4.2 Selecting data for training and testing

Since author-level predictions are computed from sets of text-level predictions, only politicians who have given at least twenty-one speeches (of at least 200 words length) are included in the training–testing rounds. In order for the basic data sets (for

Table 3 Number of individuals who have given at least twenty-one 200-word speeches, in the two classes for each classification task; the total number of such speeches given by those individuals; the number of such speeches given by the seventy-seven most productive individuals

Task	Class	Persons	Speeches	By top 77	Class	Persons	Speeches	By top 77
FM	F	208	26,071	18,533	M	256	36,952	24,574
LR	L	229	30,493	20,796	R	235	32,511	22,419
OY	O	190	21,504	15,611	Y	182	27,450	21,235
FM[L]	F	111	13,277	12,139	M	118	17,132	15,549
FM[R]	F	97	12,691	12,079	M	138	19,820	17,002
FM[O]	F	77	7,600	7,600	M	113	13,904	12,574
FM[Y]	F	83	10,407	10,242	M	99	17,043	16,348
LR[F]	L	111	13,361	12,139	R	97	12,691	12,079
LR[M]	L	118	17,132	15,549	R	138	19,820	17,002
LR[O]	L	90	9,396	9,072	R	100	12,089	11,365
LR[Y]	L	89	13,954	13,608	R	93	13,496	13,000
OY[F]	O	77	7,600	7,600	Y	83	10,407	10,242
OY[M]	O	113	13,904	12,574	Y	99	17,043	16,348
OY[L]	O	90	9,396	9,072	Y	89	13,954	13,608
OY[R]	O	100	12,089	11,365	Y	93	13,496	13,000

(A few speeches without party label, due to the fact that a member of parliament had left his/her party, are included in the FM data sets, but not for the LR tasks.)

training and testing) for each binary classification task to be balanced, they should represent the same number of politicians for each of the two classes. One way to achieve this would be to include all of the individuals of the class with the lowest number of politicians and the same number of individuals from the contrary class. Table 3 shows the number of individuals available, according to this criterion, for each task. However, for the sake of keeping the conditions as uniform as possible, the data sets represent the same number of individuals for each class for all classification tasks. Since the size of the smallest relevant class, that of older female politicians, is seventy-seven, each author set for the fifteen tasks in Table 3 comprises 154 individuals. The most productive authors are included in these sets.

As the focus of this study is on author-level classification of personal traits, the performance assessment should not reward models fitted to the idiosyncracies of individual speakers, such as personal 'keywords'. Test data are thus taken from unseen authors, i.e. authors who do not provide any of the training data. The author sets are used for a kind of tenfold cross-validation: in each training–testing round, one-tenth of the authors are reserved for testing, and the other ones provide training data. The training of the classifiers will in each case be based on 12,000 documents (delivered by the training set authors) equally divided between the two classes. All documents by authors belonging to the test set will be used for testing. An overview of the setup is given in Fig. 1.

To be more specific, for each of the fifteen binary classification tasks, tenfold cross-validation was performed (ten times) as follows: each author set, A, representing the $77+77$ authors, was partitioned into ten subsets, $A_1, \ldots, A_{10}$. In order to keep each subset roughly representative in terms of its members' productivity, the ten most productive authors in each class were randomly partitioned over the subsets. This procedure was then iterated for each category for each group of ten authors in order of falling productivity, giving us ten subsets of authors with seven or eight members of each class. Each author subset is thus balanced ($7+7$ or $8+8$) with respect to class membership and—in this particular way—with respect to productivity.

The training data for fold n ($1 \leq n \leq 10$) were compiled in such a way that they, for each class,

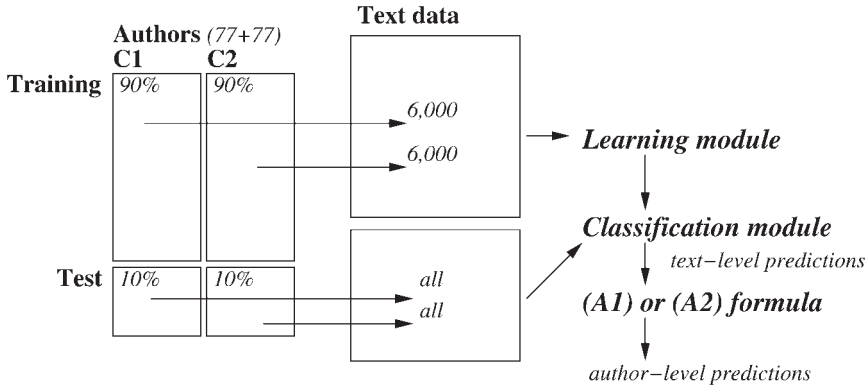

Fig. 1. The setup for each round of training and testing, C1 and C2 being the two classes involved. The learning module makes a selection of features based on the full collection of training data, maps each document to a feature vector, and calls the SVM learner to build the SVM model from these vectors. The classification module maps a document to a vector and applies the SVM model to that vector, thus retrieving the model's prediction on class membership

would comprise 6,000 texts by an author in $A - A_n$ (i.e. 12,000 200-word segments or 2.4 million words). These texts were randomly selected according to a scheme that would distribute them as evenly as possible over the training set of authors. This was achieved by arranging the authors in $A - A_n$ in a sequence, and going through that sequence iteratively, randomly selecting one speech for each author who had not—at that iterative cycle—run out of speeches. All documents by the test set authors, i.e. every at least 200-word speech given by an unseen politician (i.e. one in A_n), were used for testing.

4.3 Computing author-level predictions

There are, of course, many different ways in which we can compute a prediction of an author's membership in a class from a set of predictions concerning a number of documents by that author. This study explores two very simple schemes for doing this. Both of them involve a single text classifier and presuppose that there is a one-to-one correspondence between author classes (e.g. left-wing person) and text classes (e.g. text authored by a left-wing person).

(A1) The first formula is *majority voting*, counting each text-level prediction by the SVM as a yes or a no vote for the author-level decision.

(A2) The second way of deriving an author-level prediction is from *the sign of the mean value of the decision function values* assigned by the SVM to the texts in the set.

As can be expected, the number of texts we use to compute an author prediction, k, has substantial impact on accuracy rates. As we will see (cf. Figs 2 and 3), the (A2) formula is in general better than (A1). An assessment of the performance of the author-level classification must—given the sizes of the test data sets—be based on a large sample of combinations of the relevant texts, as the full set of relevant combinations would be too large to explore.[1] The performance scores for the author-level classification were based on the same data and classifiers as the text-level training and testing. Accuracy and recall rates for each author-level classification task and feature scheme (as in Table 5) were thus computed from 3,000,000 author-level predictions for each author, 300,000 from each of the ten (tenfold) cross-validation rounds, each prediction derived from a combination of k texts. Each combination was selected by simple random sampling. Scores for all classification tasks and all feature selection schemes will be reported for the mean decision function value formula (A2) with $k = 21$, see

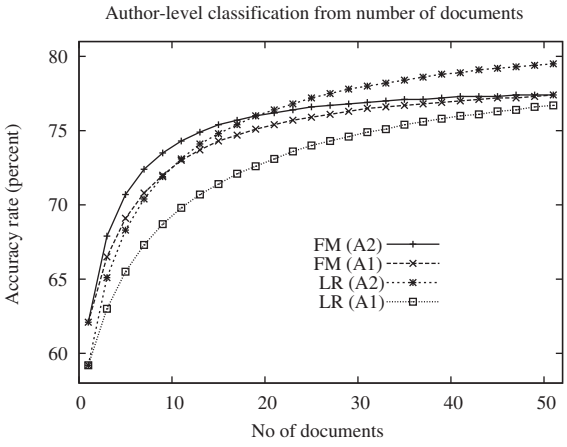

Fig. 2 Author-level predictions derived by majority vote (A1) or mean decision function value (A2)—which generally gives higher accuracy rates—from k text predictions. Two basic classification tasks: FM and LR. (Also showing OY classification would clutter the image, but see Fig. 3.) Feature scheme: *lx5k*. The author-level accuracy rates were computed from 300,000 simple random samples of k texts for each author and each of the ten classifiers built in the cross-validation rounds. (Scores were computed for odd values of k.)

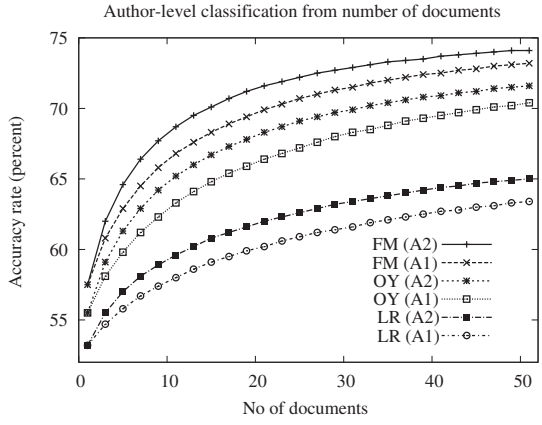

Fig. 3 Like Fig. 2, but with feature scheme *vb2k* and the three basic classification tasks: FM, LR, and OY

Table 5. The decision to compute the performance scores in this way—with 300,000 iterations for each person-classifier coupling—was motivated by the observation that repeating the full procedure gave very stable results.[2]

5 Classifiability: Results and Discussion

Since the SVM models were evaluated by tenfold cross-validation performed ten times, as described in Section 4.2, 100 different SVM models were trained for each classification task and feature selection scheme. Each document by the 154 authors delivering data was a test document for ten classifiers. Table 4 shows the mean recall rates for both classes for the various classification situations.

As we saw in Section 2, the performance of classifiers of the present kind is most often quantified in terms of accuracy rates. The import of these scores depend on the proportion of positive and negative examples in the test data, as a baseline accuracy rate is defined by assigning to each example the most common category. Since the text-level test data are not balanced here, recall rates have been preferred over accuracy rates to quantify the performance. The recall rate corresponds to the accuracy for the subset of data belonging to the target category. In relation to a binary classification task, the mean value of the two recall rates represents a normalized accuracy score. As the number of examples involved is reported in the recall tables, accuracy rates can be computed. Furthermore, recall rates have the advantage of revealing performance asymmetries over categories in the behaviour of the classifiers.

Examples of the performance of the two author-level prediction schemes proposed in Section 4.3, (A1) and (A2)—for two of the feature selection schemes, *lx5k* and *vb2k* and for two or three of the basic classification tasks—are shown in Figs 2 and 3. We see that the author-level accuracy rates improve as the number of texts used to compute the predictions increases. Moreover, the mean decision function value formula (A2) in general gives better accuracy rates than majority vote (A1). Using (A2) to make author-level predictions from twenty-one text predictions, as described in Section 4.3, gives us the recall rates in Table 5.

Tables 4 and 5 show that all classification task and feature scheme combinations give accuracy rates above the baseline level. The highest scores are for the *all5k* scheme and the political (LR) classification tasks, and the lowest ones are for

Table 4 Text-level recall rates (percent) for each target category obtained for the setup in Section 4.2 (means over ten author-based tenfold cross-validations)

Task	*n*		*all5k*		*lx5k*		*lx2k*		*vb2k*		*fu1k*	
	F	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
FM	18,533	24,574	59.7	62.3	60.2	62.4	59.7	62.0	55.8	57.7	58.3	57.5
FM[L]	12,139	15,549	57.1	61.3	57.6	61.5	57.0	61.1	54.7	57.4	57.5	55.5
FM[R]	12,079	17,002	58.4	65.4	58.9	65.7	58.5	65.7	55.6	60.1	56.4	59.0
FM[O]	7,600	12,574	61.1	67.8	61.6	67.8	61.3	67.4	58.3	61.3	60.7	58.3
FM[Y]	10,242	16,348	58.9	62.3	58.4	63.0	58.1	62.9	56.5	56.7	55.6	59.5
	L	R	L	R	L	R	L	R	L	R	L	R
LR	20,796	22,419	65.0	66.1	59.0	59.0	58.8	58.4	52.1	54.5	54.0	53.5
LR[F]	12,139	12,079	63.4	64.8	59.0	59.2	58.8	58.9	54.1	53.3	54.5	51.1
LR[M]	15,549	17,002	65.2	66.1	59.5	59.3	59.3	59.4	53.4	54.4	55.2	54.0
LR[O]	9,072	11,365	64.2	66.3	60.5	60.7	60.0	60.1	54.7	55.7	57.1	56.5
LR[Y]	13,608	13,000	65.2	67.6	59.3	60.6	58.8	59.2	53.5	54.3	54.2	53.2
	O	Y	O	Y	O	Y	O	Y	O	Y	O	Y
OY	15,610	21,235	58.5	60.4	58.6	60.0	58.4	59.8	54.0	56.4	54.1	54.6
OY[F]	7,600	10,242	60.9	60.9	60.3	59.7	59.7	59.3	55.6	56.4	57.8	54.3
OY[M]	12,574	16,348	58.1	62.1	58.8	62.3	58.5	62.1	56.2	56.5	53.3	58.0
OY[L]	9,072	13,608	58.3	59.0	58.8	58.4	58.9	58.3	53.8	53.0	54.7	57.7
OY[R]	11,365	13,000	60.7	64.7	60.8	64.1	60.7	63.3	57.8	60.3	56.2	58.3

Fifteen classification tasks, as defined in Table 1. Five feature selection schemes. The *n* columns give the number of documents (i.e. 200-word initial segments of speeches) delivered by the seventy-seven authors representing each class.

Table 5 Author-level recall rates (percent)

Task	*all5k*		*lx5k*		*lx2k*		*vb2k*		*fu1k*	
	F	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**
FM	76.7	75.6	76.9	75.5	76.3	75.0	72.0	71.1	75.7	71.3
FM[L]	70.1	73.5	71.3	74.3	71.1	73.2	70.5	70.9	72.6	68.0
FM[R]	80.1	78.5	80.5	79.7	79.1	79.3	77.0	73.1	71.0	75.3
FM[O]	76.7	83.6	78.3	84.2	77.8	83.6	78.8	76.0	77.0	77.9
FM[Y]	72.7	72.4	72.3	73.0	71.2	72.9	73.3	63.6	66.9	72.4
	L	R	L	R	L	R	L	R	L	R
LR	85.3	91.2	72.8	80.1	71.5	78.9	60.3	63.8	61.4	61.6
LR[F]	82.0	93.2	74.9	84.3	74.2	83.6	65.3	70.0	64.4	61.5
LR[M]	85.8	89.9	75.6	79.8	74.4	79.3	62.4	65.1	67.5	63.4
LR[O]	83.5	90.2	77.1	82.7	75.6	81.2	64.6	74.9	70.8	72.9
LR[Y]	85.8	93.0	77.2	83.6	75.7	81.6	63.3	66.0	64.0	62.0
	O	Y	O	Y	O	Y	O	Y	O	Y
OY	74.8	74.2	75.2	74.0	74.6	72.8	68.6	68.0	65.6	67.2
OY[F]	77.5	77.7	78.0	75.0	76.9	74.1	74.0	69.0	75.0	65.8
OY[M]	77.4	75.6	78.4	77.2	78.1	77.0	73.8	67.7	67.9	75.6
OY[L]	74.6	70.6	76.3	70.2	75.7	69.0	63.6	63.2	71.0	70.7
OY[R]	78.6	79.2	79.2	78.6	78.9	77.2	80.7	74.6	70.6	73.3

The author-level predictions are based on the mean decision function value formula (A2) (see Section 4.3) for twenty-one text segments (4,200 words) and the corresponding text-level predictions, whose performance scores are given in Table 4.

the same task and the *vb2k* and *fu1k* schemes. The gender (FM) and age (OY) classifiers perform more evenly. Recall rates are more or less asymmetric over the two classes.

When we compare the performance scores for the general populations and the restricted cohorts, we see that gender differentiation is stronger in the older group (FM[O]) than among younger politicians (FM[Y]). This is clearly attested by the recall rates across the board. A similar tendency can be seen for the right-wing politicians (FM[R]) in comparison to the left-wing group (FM[L]). The difference is strongly reflected in the relevant recall rates. These findings suggest that left-wing and younger politicians are more gender-equal vocabulary-wise than the right-wing and the older group, respectively.

Another tendency is that age prediction in most cases gives higher recall scores for the right-wing cohort (OY[R]) than for the left-wing group (OY[L]). Furthermore, political affiliation classification gives higher accuracy rates for the restricted cohorts for all features selection schemes except *all5k*. This might indicate that many word forms are more informative about political affiliation within the restricted cohorts—LR[F], LR[M], LR[O], and LR[Y]—than for the general group. This effect may also be related to the fact that the data for the restricted cohorts are derived from less productive speakers than those for the general LR task.

We find the highest recall rates for the *all5k* scheme—where proper nouns are included among the features—and the five LR tasks, for which the *lx5k* scheme performs considerably worse. For the ten FM and OY classification tasks, the predictive power of the *all5k* scheme is much lower. So, proper nouns are of considerable value as features in political affiliation classification. Part of the explanation for this is that members of parliament often address and talk about politicians of the opposite camp. In contrast, leaving proper nouns out does not make much of a difference in gender and age prediction, as witnessed by the small differences in performance between the *all5k* and *lx5k* feature schemes for the ten FM and OY classification tasks.

With few exceptions, the *lx2k* scheme gives lower recall rates than the *lx5k* scheme. The differences are, however, quite small. The verb only scheme, *vb2k*, generally performs considerable worse than the *lx2k* scheme. This is more pronounced for the LR tasks, for which the *vb2k* scheme gives the lowest recall rates, than in the FM and OY cases. We also see that the OY[L] scores for *vb2k* are exceptionally low.

As can be expected, the recall rates for the function word scheme, *fu1k*, are much lower than those for the *lx2k* scheme. When it comes to the political affiliation classification tasks, the scores suggest that the predictive power of the *fu1k* scheme is much stronger for the older cohort (LR[O]) and somewhat stronger for the male cohort (LR[M]) than for the opposite restricted tasks. For author-level age classification, the *fu1k* scheme generally performs better for the restricted OY tasks. When we compare *fu1k* and *vb2k*, we see that their relative performance varies for the different tasks. The *vb2k* and *fu1k* schemes are markedly better for the older cohorts both as regards gender (FM[O]) and political (LR[O]) author-level classification than for the other restricted cohorts and for the general groups.

A detailed analysis of key words as retrieved by the feature extraction would be the topic of another article, but a few examples could be of interest here. Table 6 lists the top ten features as ranked by IG, for one fold of one (out of ten) of the tenfold cross-validation rounds for the unrestricted classification tasks, i.e. FM, LR, and OY, and the *lx5k* feature scheme (which in this regard is equal to *lx2k*). Features marked with an asterisk are among the top 100 for all 100 training sets. Many of these features seem to point in the direction of 'content themes', such as 'women–men–children' for female speakers—thus replicating findings from studies mentioned in Section 2—and 'education' for the younger group. Some more topic-neutral words—adverbs, notably—are also top-ranked for gender and age classification.

6 Conclusions

The aim of the present study was to explore the predictability of personal traits in Swedish

Table 6 Examples of top-ranked features (word forms) for one training data set, the *lx5k*/*lx2k* feature scheme, and the three general classification tasks: target class, frequency for the target class, F_C, frequency for the opposite class, F_O, and the IG value (multiplied by 1000)

Target	Feature	F_C	F_O	IG
The (general) FM task				
F	kvinnor*, 'women'	536	211	9.4
F	också*, 'also'	3669	3119	6.2
F	barn*, 'child(ren)'	597	318	5.6
F	dag, 'day'	1821	1405	4.4
M	möjligen*, 'possibly'	144	36	4.2
F	behöver*, 'need(s)' (v.)	907	627	3.5
F	otroligt, 'incredible/-ly'	144	43	3.5
F	män*, 'men'	243	106	3.4
F	barnen*, 'the children'	231	100	3.3
F	kulturministern, 'the minister of culture'	77	13	3.1
The (general) LR task				
L	borgerliga*, 'right-wing'	524	177	11.4
L	arbetsmarknadsministern*, 'the minister of labour'	137	16	6.7
R	kristdemokrater*, 'Christian Democrats'	106	13	5.0
R	socialdemokratiska*, 'social democratic'	337	165	3.8
R	interpellanten*, 'the interpellator'	125	31	3.7
L	ministern, 'the minister'	421	231	3.6
R	allians regeringen*, 'the alliance government'	143	42	3.6
L	moderatledda, 'Moderate-led'	46	2	3.0
L	a-kassan*, 'the unemployment insurance'	116	36	2.7
L	arbetslösa*, 'unemployed'	197	89	2.6
The (general) YG task				
Y	skolan*, 'the school'	334	105	7.8
Y	elever, 'pupils'	186	44	5.8
Y	förskolan*, 'the preschool'	93	16	3.6
Y	eleverna, 'the pupils'	111	25	3.6
Y	skolans, 'the school's'	56	6	2.8
Y	poliser, 'policemen'	103	28	2.8
Y	lärare, 'teacher(s)'	144	52	2.7
Y	polisen, 'the police(man)'	169	71	2.5
O	naturligtvis*, 'of course'	888	651	2.5
Y	jobb, 'job(s)'	456	289	2.4

Features marked '*' are among the top 100 for all 100 training sets.

politicians, based on the wording of speeches they have delivered in the Swedish parliament. The study was concerned with three binary trait distinctions: gender (female versus male), age (born 1953 or earlier versus born 1959 or later), and political affiliation (left-wing versus right-wing). Classification was performed by means of SVM models trained on and targeting 200-word stretches of text. Simple binary bag-of-word-forms vectors were used to represent the documents. One of the objectives was to compare different cohorts of politicians, defined in terms of one of the basic personal traits, as regards the performance of classifiers concerned with another trait distinction. Classification concerning each basic distinction was thus applied to general populations of politicians and to cohorts restricted to the basic classes of the other two classification tasks. Furthermore, the setup included five feature selection schemes designed to elucidate to what extent different personal traits affect different layers of the vocabulary. The assessment of the classifiers' performance relied on data from authors unseen in the training data.

The basic text-level SVM classifiers gave accuracy rates well above the baseline for each cohort and feature selection scheme. We also saw that much higher accuracy rates can be achieved by computing author-level predictions from SVM decision function values for a set of basic documents. The classifiers' performance scores consequently showed that membership in the six basic classes to a considerable degree is reflected in the speakers' choice of words. The best author-level accuracy rates (i.e. means of the paired recall rates in Table 5) obtained from twenty-one 200-word text segments (i.e. 4,200 words) were 81.2% for gender prediction (FM[O] and *lx5k*), 89.4% for political affiliation prediction (LR[Y] and *all5k*), and 78.9% for age prediction (OY[R] and *lx5k*).

The most striking cohort-related tendencies that the study brought to light are the following (with accuracy rates obtained from twenty-one 200-word text segments and the *lx5k* feature selection scheme, which for the classification tasks involved gave the best performance):

(1) The accuracy rates for gender prediction are considerably higher for the older cohort (81.2%) than for the younger cohort (72.7%).
(2) The accuracy rates for gender prediction are considerably higher for the right-wing

cohort (80.1%) than for the left-wing cohort (72.8%).

(3) The accuracy rates for age prediction are considerably higher for the right-wing cohort (78.9%) than for the left-wing cohort (73.3%).

The differences in (1) may correspond to the progress of gender equality, which for the past 40 years or so has been one of the central concerns for Swedish law making. We may interpret (2) as reflecting a higher degree of vocabulary-related gender equality among left-wing speakers. This, in turn, may be the effect of a stronger general emphasis on gender equality in left-wing politics than in right-wing thinking. The tendency in (3) may be connected to the efforts of the right-wing parties of moving towards the political center, i.e. as highlighted by the 'Nya Moderaterna' ('New Moderates') branding of the Moderate Party. This might have left older right-wing politicians with a more markedly old-fashioned vocabulary, whereas the rhetoric of the left-wing parties has stressed their traditional roots in the labour and green politics movements.

The performance of the feature selection schemes revealed that unrestricted selection of graph words, which admitted proper nouns among the features, gave higher accuracy rates for political affiliation classification than for the other classification tasks. This tendency may have been strengthened by the choice to use only introductory portions of the speeches—where addressees and party affiliations are often explicitly named—as basic document data. On the other hand, when features were restricted to verbs or function words (known from another corpus), the classifiers gave the highest accuracy rates for gender prediction, and the lowest ones for political affiliation classification. These results suggest that the more dynamic layers of the vocabulary give more information about political affiliation than about gender and age, whereas the absence or presence of forms from the more closed word classes has more to tell us about gender and age than about political affiliation.

In view of the wide range of interesting data, encompassing e.g. different languages, places, times, genres, institutional contexts, and ways of extracting basic documents, and the multitude of possibilities for classifier design, current knowledge of text- and author-level classification concerning personal traits must be considered fragmentary. An aspect of this is that the performance scores reported here can only be compared with results obtained for data which in several respects are crucially different from the present data sets. These were drawn from professionally edited transcripts of speeches given in a impersonal and formal setting, where perspicuous display of conventional gender traits are avoided, as are references to the politicians' private relations and activities. In contrast, many earlier studies on personal trait prediction have been concerned with genres encouraging more personal modes of expression. A technically important parameter is the length of the basic documents, as it is possible for longer documents to reflect relative frequencies for a larger number of forms and in a more representative way than a short document can do. So, it is likely that feature weighting based on relative frequency is more useful the longer the documents to be classified are, whereas binary weighting would be better for short documents, as in the present study.

The classification methods that have been applied here are based on state-of-the-art software and simple feature engineering. It is plausible that the use of features based on more sophisticated grammatical and semantic analysis would improve classifier performance and reveal other interesting connections between personal traits and linguistic features. For instance, political affiliation classification invite the application of methods involving opinion mining and extraction of political affiliation and addressee reference phrases.

Further studies are needed to decide whether the tendencies revealed by the present study are of a more general validity. Experiments involving other feature generation schemes and classifier frameworks, or set up in a different way, but performed on the same data sets, could, for instance, strengthen or weaken the support for the cohort-related conclusions, as reported in (1)–(3). Another obvious path for further research would be to perform studies in the same style on legislative

discourse in other languages or to extend our attention to other genres.

Acknowledgements

The author wishes to thank Joakim Nivre for detailed comments on a previous version of this article and Bengt Dahlqvist, Marco Kuhlmann, and Beáta Megyesi for discussions on an earlier presentation.

References

Argamon, S., Goulain, J.-B., Horton, R., and Olsen, M. (2009). Vive la Différence! Text mining gender difference in French literature. *Digital Humanities Quarterly*, **3**(2). http://www.digitalhumanities.org/dhq/vol/3/2/000042.html.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2007). Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday*, **12**(9). http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2003/1878.

Boulis, C. and Ostendorf, M. (2005). A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. New Brunswick, NJ: Association for Computational Linguistics (ACL), pp. 435–442.

Corney, M., de Vel, O., Anderson, A., and Mohay, G. (2002). Gender-preferential Text Mining of E-mail Discourse. *Proceedings of the 18th Annual Computer Security Applications Conference*. Los Alamitos: IEEE Computer Society, pp. 282–289.

Diermeier, D., Godbout, J.-F., Yu, B., and Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science*, **42**(1): 31–55.

Ejerhed, E., Källgren, G., and Brodda, B. (2006). *Stockholm Umeå Corpus Version 2.0, SUC 2.0* Department of Linguistics, Stockholm University.

Greevy, E. P. and Smeaton, A. F. (2004). Text Categorisation of Racist Texts Using a Support Vector Machine. *Le poids des mots: Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT vol.1)*. Presses universitaires de Louvain.

Hota, S., Argamon, S., and Chung, R. (2006). Gender in Shakespeare: Automatic Stylistics Gender Classification Using Syntactic, Lexical, and Lemma Features. *Chicago Colloquium on Digital Humanities and Computer Science*.

Jiang, M. and Argamon, S. (2008). Political Leaning Categorization by Exploring Subjectivities in Political Blogs. *Proceedings of the Fourth International Conference on Data Mining (DMIN 2008)*. New York: ACM, pp. 725–26.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning*. London: Springer, pp. 137–42.

Joachims, T. (1999). Making Large-scale Support Vector Machine Learning Practical. In Schölkopf, B., Burges, C., and Smola, A. (eds), *Advances in Kernel Methods – Support Vector Learning*. Cambridge, MA: MIT Press, pp. 169–84.

Kim, L. and Daelemans, W. (2008). Using Syntactic Features to Predict Author Personality from Text. *Proceedings of Digital Humanities 2008*. Oulu, Finland: University of Oulu, pp. 146–49.

Kobayashi, D., Matsumura, N., and Ishizuka, M. (2007). Automatic Estimation of Bloggers' Gender. *Proceedings of International Conference on Weblogs and Social Media*. Boulder: Omnipress.

Komiya, K., Igarashi, C., Shibahara, K., Fujimoto, K., Tajima, Y., and Kotani, Y. (2009). Generating a set of rules to determine the gender of a speaker of a Japanese sentence. *WSEAS Transactions on Communications*, **8**: 112–21.

Koppel, M., Akiva, N., Alshech, E., and Bar, K. (2009). Automatically classifying documents by ideological and organizational affiliation. *Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics*. Piscataway: IEEE Press, pp. 176–78.

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

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., and Can, F. (2006). Chat Mining for Gender Prediction. *Advances in Information Systems (Proceedings of the 4th International Conference on Advances in Information Systems, ADVIS 2006)*. Berlin, Heidelberg, and New York: Springer, pp. 274–83.

Mukherjee, A. and Liu, B. (2010). Improving Gender Classification of Blog Authors. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Stroudsburg: Association for Computational Linguistics, pp. 207–17.

Nowson, S. (2006). The Language of Weblogs: A Study of Genre and Individual Differences. http://nowson.com/papers/NowsonThesis06.pdf (Accessed 29 June 2011).

Nowson, S. and Oberlander, J. (2006). The Identity of Bloggers: Openness and Gender in Personal Weblogs. *Computational Approaches to Analyzing Weblogs: Papers from the AAAI Spring Symposium*. Menlo Park: The AAAI Press, pp. 163–167.

Opsomer, R., Knoth, P., van Polen, F., Trapman, J., and Wiering, M. (2008). Categorizing Children Automated Text Classification of CHILDES Files. *Proceedings 20th Belgian-Netherlands Conference on Artificial Intelligence*. Enschede: Universiteit Twente, pp. 209–15.

Otterbacher, J. (2010). Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. New York: ACM, pp. 369–78.

Pion, S. and Hamel, L. (2007). Comparing the Results of Support Vector Machines with Traditional Data Mining Algorithms. *Proceedings of the 2007 International Conference on Data Mining (DMIN 2007)*. Athens, GA: Computer Science Research, Education, and Applications Press, pp. 79–83.

Sabin, R. E., Goodwin, K. A., Goldstein-Stewart, J., and Pereira, J. A. (2008). Gender Differences across Correlated Corpora: Preliminary Results. *Proceedings of the Twenty-First International FLAIRS Conference*. Menlo Park: AAAI Press, pp. 207–12.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. (2006). Effects of Age and Gender on Blogging. *Computational Approaches to Analyzing Weblogs: Papers from the AAAI Spring Symposium*. Menlo Park: The AAAI Press, pp. 199–205.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*. San Francisco: Morgan Kaufmann Publishers Inc., pp. 412–420.

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

Zhang, Y., Dang, Y., and Chen, H. (2009). Gender Difference Analysis of Political Web Forums: an Experiment on an International Islamic Women's Forum. *Proceedings of the 2009 IEEE International Conference on Intelligence and Security informatics*. Piscataway: IEEE Press, pp. 61–4.

Zhang, Y., Dang, Y., and Chen, H. (2011). Gender classification for web forums. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, **41**(4): 668–77.

Notes

1 The number of different combinations of texts—which is given by the binomial coefficient $\binom{n}{k}$, where n is the cardinality of the full set of texts—varied between $\binom{21}{21} = 1$ and $\binom{925}{21} > 3.0 \times 10^{42}$, when $k = 21$, as 925 is the largest number of documents available for one individual.

2 Repeating the author-level recall score computation procedure three times gave differences of at most 0.1 percentage point for at most four of the percentage values rounded to one decimal, as in Table 5. In those cases, the score reported was reached in two of the three rounds.