Authorship Attribution and Verification with Many Authors and Limited Data

Kim Luyckx and **Walter Daelemans**
CNTS Language Technology Group
University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
{kim.luyckx,walter.daelemans}@ua.ac.be

Abstract

Most studies in statistical or machine learning based authorship attribution focus on two or a few authors. This leads to an overestimation of the importance of the features extracted from the training data and found to be discriminating for these small sets of authors. Most studies also use sizes of training data that are unrealistic for situations in which stylometry is applied (e.g., forensics), and thereby overestimate the accuracy of their approach in these situations. A more realistic interpretation of the task is as an *authorship verification* problem that we approximate by pooling data from many different authors as negative examples. In this paper, we show, on the basis of a new corpus with 145 authors, what the effect is of many authors on feature selection and learning, and show robustness of a memory-based learning approach in doing authorship attribution and verification with many authors and limited training data when compared to eager learning methods such as SVMs and maximum entropy learning.

1 Introduction

In traditional studies on authorship attribution, the focus is on small sets of authors. Trying to classify an unseen text as being written by one of two or of a few authors is a relatively simple task, which in most cases can be solved with high reliability and accuracies over 95%. An early statistical study by Mosteller and Wallace (1964) adopted distributions of function words as a discriminating feature to settle the disputed authorship of the Federalist Papers between three candidate authors (Alexander Hamilton, James Madison, and John Jay). The advantage of distributions of function words and syntactic features is that they are not under the author's conscious control, and therefore provide good clues for authorship (Holmes, 1994). Frequencies of rewrite rules (Baayen et al., 1996), *n*-grams of syntactic labels from partial parsing (Hirst and Feiguina, 2007), *n*-grams of parts-of-speech (Diederich et al., 2000), function words (Miranda García and Calle Martín, 2007), and functional lexical features (Argamon et al., 2007) have all been claimed to be reliable markers of style. There is of course a difference between claims about types of features and claims about individual features of that type. E.g., it may be correct to claim that distributions of function words are important markers of author identity, but the distribution of a particular function word, while useful to distinguish between one particular pair of authors, may be irrelevant when comparing another pair of authors.

The field of authorship attribution is however dominated by studies potentially overestimating the importance of these specific predictive features in experiments discriminating between only two or a few authors. Taking into account a larger set of authors allows the computation of the degree of variability encountered in text on a single topic of different (types of) features. Recently, research has started to focus on authorship attribution on larger sets of authors: 8 (Van Halteren, 2005), 20 (Argamon et al., 2003), 114 (Madigan et al., 2005), or up

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 513–520
Manchester, August 2008

to thousands of authors (Koppel et al., 2006) (see Section 5).

A second problem in traditional studies are the unrealistic sizes of training data, which also makes the task considerably easier. Researchers tend to use over 10,000 words per author (Argamon et al., 2007; Burrows, 2007; Gamon, 2004; Hirst and Feiguina, 2007; Madigan et al., 2005; Stamatatos, 2007), which is regarded to be 'a reliable minimum for an authorial set' (Burrows, 2007). When no long texts are available, for example in poems (Coyotl-Morales et al., 2006) or student essays (Van Halteren, 2005), a large number of short texts is selected for training for each author. One of the few studies focusing on small texts is Feiguina and Hirst (2007), but they select hundreds of these short texts (here 100, 200 or 500 words). The accuracy of any of these studies with unrealistic sizes of training data is overestimated when compared to realistic situations. When only limited data is available for a specific author, the author attribution task becomes much more difficult. In forensics, where often only one small text per candidate author is available, traditional approaches are less reliable than expected from reported results.

In this paper, we present a more realistic interpretation of the authorship attribution task, viz. as a problem of *authorship verification*. This is a much more natural task, since the group of potential authors for a document is essentially *unknown*. Forensic experts not only want to identify the author given a small set of suspects, they also want to make sure the author is not someone else not under investigation. They often deal with short e-mails or letters and have only limited data available. The central question in authorship verification is *Did candidate author x write the document?* Of the three basic approaches to authorship verification - also including a one-class learning approach (Koppel et al., 2007) - we selected a *one vs. all* approach. This approach is similar to the one investigated by Argamon et al. (2003), which allows for a better comparison of results. With only few positive instances and a large number of negative instances to learn from, we are dealing with highly skewed class distributions.

We show, on the basis of a new corpus with 145 authors, what the effect is of many authors on feature selection and learning, and show robustness of a memory-based learning approach in doing authorship attribution and verification with many authors and limited training data when compared to eager learning methods such as SVMs and maximum entropy learning. As far as feature selection is concerned, we find that similar *types* of features tend to work well for small and large sets of authors, but that no generalisations can be made about *individual* features. Classification accuracy is clearly overestimated in authorship attribution with few authors. Experiments in authorship verification with a *one vs. all* approach reveal that machine learning methods are able to correctly classify up to 56% of the positive instances in test data.

For our experiments, we use the *Personae* corpus, a collection of student essays by 145 authors (see Section 2). Most studies in stylometry focus on English, whereas our focus is on Dutch written language. Nevertheless, the techniques used are transferable to other languages.

2 Corpus

The 200,000-word *Personae* corpus[1] used in this study consists of 145 student (BA level) essays of about 1400 words about a documentary on Artificial Life, thereby keeping markers of genre, register, topic, age, and education level relatively constant. These essays contain a factual description of the documentary and the students' opinion about it. The task was voluntary and students producing an essay were rewarded with two cinema tickets. The students also took an online Myers-Briggs Type Indicator (MBTI) (Briggs Myers and Myers, 1980) test and submitted their profile, the text and some user information via a website. All students released the copyright of their text and explicitly allowed the use of their text and associated personality profile for research, which makes it possible to distribute the corpus. The corpus cannot only be used for authorship attribution and verification experiments, but also for personality prediction. More information about the motivation behind the corpus and results from exploratory experiments in personality prediction can be found in Luyckx & Daelemans (2008).

3 Methodology

We approach authorship attribution and verification as automatic text categorization tasks that label documents according to a set of predefined categories (Sebastiani, 2002, 3). Like in most text cat-

[1] The *Personae* corpus can be downloaded from http://www.cnts.ua.ac.be/~kim/Personae.html

egorization systems, we take a two-step approach in which our system (i) achieves automatic selection of features that have high predictive value for the categories to be learned (see Section 3.1), and (ii) uses machine learning algorithms to learn to categorize new documents by using the features selected in the first step (see Section 3.2).

3.1 Feature Extraction

Syntactic features have been proposed as more reliable style markers than for example token-level features since they are not under the conscious control of the author (Baayen et al., 1996; Argamon et al., 2007). To allow the selection of linguistic features rather than (*n*-grams of) terms, robust and accurate text analysis tools such as lemmatizers, part of speech taggers, chunkers etc., are needed. We use the Memory-Based Shallow Parser (MBSP) (Daelemans and van den Bosch, 2005), which gives an incomplete parse of the input text, to extract reliable syntactic features. MBSP tokenizes the input, performs a part-of-speech analysis, looks for noun phrase, verb phrase and other phrase chunks and detects subject and object of the sentence and a number of other grammatical relations.

Word or part-of-speech (*n*-grams) occurring more often than expected with either of the categories are extracted automatically for every document. We use the χ^2 metric (see Figure 1), which calculates the expected and observed frequency for every item in every category, to identify features that are able to discriminate between the categories under investigation.

$$\boxed{\chi^2 = \sum_{i=1}^{k} \frac{(\chi_i - \mu_i)^2}{\sigma_i}}$$

Figure 1: Chi-square formula

Distributions of *n*-grams of lexical features (*lex*) are represented numerically in the feature vectors, as well as of *n*-grams of both fine-grained (*pos*) and coarse-grained parts-of-speech (*cgp*). The most predictive function words are present in the *fwd* feature set. For all of these features, the χ^2 value is calculated.

An implementation of the Flesch-Kincaid metric indicating the readability of a text, along with its components (viz., mean word and sentence length) and the type-token ratio (which indicates vocabulary richness) are also represented (*tok*).

3.2 Experimental Set-Up

This paper focuses on three topics, each with their own experimental set-up:

(a) the effect of many authors on feature selection and learning;

(b) the effect of limited data in authorship attribution;

(c) the results of authorship verification using many authors and limited data on learning.

For (a), we perform experiments in authorship attribution while gradually increasing the number of authors. First, we select a hundred random samples of 2, 5 and 10 authors in order to minimize the effect of chance, then select one random sample of 20, 50, 100 authors and finally experiment with all 145 authors (Section 5.1).

We investigate (b) by performing authorship attribution on 2 and 145 authors while gradually increasing the amount of training data, keeping test set size constant at 20% of the entire corpus. The resulting learning curve will be used to compare performance of eager and lazy learners (see Section 5.1).

The authorship verification task (c) - which is closer to a realistic situation in e.g. forensics - using limited data and many authors is approximated as a skewed binary classification task (*one vs. all*). For each of the 145 authors, we have 80% of the text in training and 20% in test. The negative class contains 80% of each of the other 144 author's training data in training and 20% in test (see Section 5.2).

All experiments for (a), (b) and (c) are performed using 5-fold cross-validation. This allows us to get a reliable indication of how well the learner will do when it is asked to make new predictions on the held-out test set. The data set is divided into five subsets containing two fragments of equal size per author. Five times one of the subsets is used as test set and the other subsets as training set.

The feature vectors that are fed into the machine learning algorithm contain the top-*n* features (*n*=50) with highest χ^2 value. Every text fragment is split into ten equal parts, each part being represented by means of a feature vector, resulting in 1450 vectors per fold (divided over training and test).

For classification, we experimented with both lazy and eager supervised learning methods. As an implementation of the lazy learning approach we used TiMBL (Tilburg Memory-Based Learner) (Daelemans et al., 2007), a supervised inductive algorithm for learning classification tasks based on the *k-nn* algorithm with various extensions for dealing with nominal features and feature relevance weighting. Memory-based learning stores feature representations of training instances in memory without abstraction and classifies new instances by matching their feature representation to all instances in memory. From these 'nearest neighbors', the class of the test item is extrapolated.

As eager learners, we selected SMO, an implementation of Support-Vector Machines (SVM) using Sequential Minimal Optimization (Platt, 1998), and Maxent, an implementation of Maximum Entropy learning (Le, 2006). SMO is embedded in the WEKA (Waikato Environment for Knowledge Analysis) software package (Witten and Frank, 1999).

Our expectation is that eager learners will tend to overgeneralize for this task when dealing with limited training data, while lazy learners, by delaying generalization over training data until the test phase, will be at an advantage when dealing with limited data. Unlike eager learners, they will not ignore - i.e. not abstract away from - the frequently occurring infrequent or untypical patterns in the training data, that will nevertheless be useful in generalization.

4 Results and Discussion

In this section, we present results of experiments concerning the three main issues of this paper (see Section 3.2 for the experimental set-up):

(a) the effect of many authors on feature selection and learning;

(b) the effect of limited data in authorship attribution;

(c) the results of authorship verification using many authors and limited data on learning.

4.1 Authorship Attribution

(a) Figure 2 shows the effect of many authors in authorship attribution experiments using memory-based learning (TiMBL) (k=1) and separate feature sets. Most authorship attribution studies focus on a small set of authors and report good results, but systematically increasing the amount of authors under investigation leads to a significant decrease in performance. In the 2-author task (100 experiments with random samples of 2 authors), we achieve an average accuracy of 96.90%, which is in line with results reported in other studies on small sets of authors. The 5-, 10- (both in 100 experiments with random samples) and 20-author tasks show a gradual decrease in performance with results up to 88%, 82% and 76% accuracy, respectively. A significant fall in accuracy comes with the 50- and 100-author attribution task, where accuracy drops below 52% for the best performing feature sets. Experiments with all 145 authors from the corpus (as a multiclass problem) show an accuracy up to 34%. Studies reporting on accuracies over 95% are clearly overestimating their performance on a small set of authors.

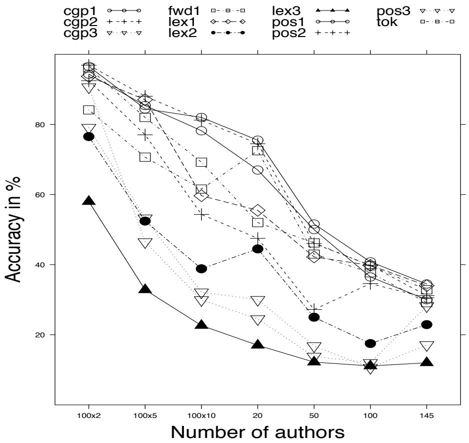

Figure 2: The effect of many authors using single feature sets

Incremental combinations of feature sets performing well in authorship attribution lead to an accuracy of almost 50% in the 145-author case, as is shown in Figure 3. This indicates that providing a more heterogeneous set of features improves the system significantly. Memory-based learning shows robustness for a large set of authors in authorship attribution.

As far as feature selection is concerned, we find that similar *types* of features tend to work well for small and large sets of authors in our corpus, but that no generalisations can be made about *individual* features towards other corpora or studies, since this is highly dependent of the specific authors selected.

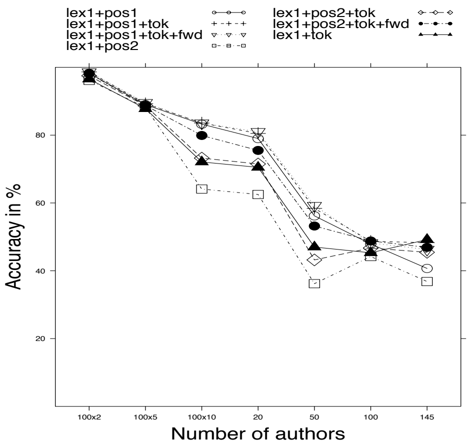

Figure 3: The effect of many authors using combinations of feature sets

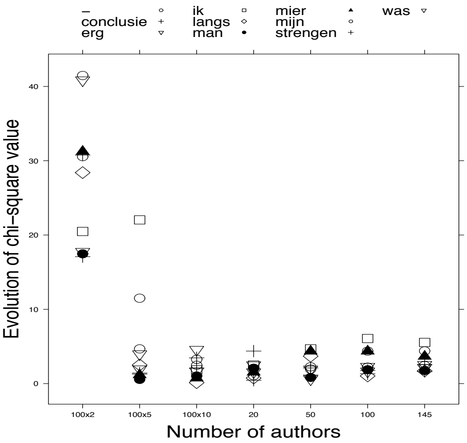

Figure 4: The effect of many authors on χ^2 value for 2-author discriminating features

Figure 4 shows the top-ten features with highest χ^2 value in one of the randomly selected 2-author samples. In 5-author cases, we see that some of these features have some discriminating power, but with the increase of the number of authors comes a decrease in importance.

(b) The effect of limited data is demonstrated by means of a learning curve. The performance of lazy learner TiMBL is compared to that of eager learners Maxent (Maximum Entropy Learning) and SMO (Support-Vector Machines) when comparing different training set sizes. Figure 5 shows the evolution of learning in authorship attribution using the *lex1* feature set. Although memory-based learning does show robustness when dealing with limited data, we cannot show a clear superiority on this aspect to the eager learning methods in this experiment. However, results are good enough to warrant continuing the experiments on authorship verification with this method.

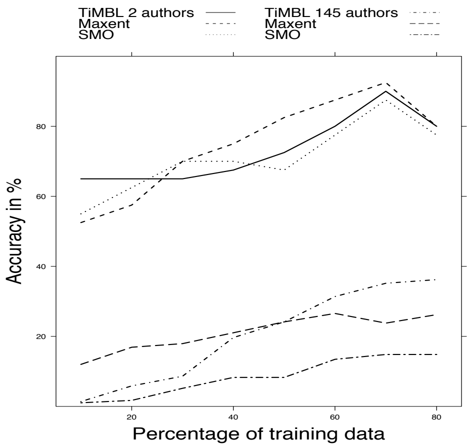

Figure 5: The effect of limited data in authorship attribution on *lex1*

4.2 Authorship Verification

(c) We now focus on a more realistic interpretation of the authorship attribution task, viz. as a authorship *verification* problem. Forensic experts want to answer both questions of authorship attribution (*Which of the n candidate authors wrote the document?*) and verification (*Did candidate author x write the document?*). They often deal with limited data like short e-mails or letters, and the amount of candidate authors is essentially unknown. With only few positive instances (of 1 author) and a large amount of negative instances (of 144 authors in our corpus), we are dealing with highly skewed class distributions.

We approximate the author verification problem by defining a binary classification task with the author fragments as positive training data, and the fragments of all the other authors as negative training data. A more elegant formulation would be as a one-class problem (providing only positive training data), but in exploratory experiments, these one-class learning approaches did not yield useful results.

We evaluate authorship verification experiments by referring to precision and recall of the positive class. Recall represents the proportional number of times an instance of the positive class has correctly been classified as positive. Precision shows the proportion of test instances predicted by the system to be positive that was correctly classified as such.

Feature set	Precision	Recall	F-score
tok	20.66%	15.93%	17.99%
fwd	37.89%	8.41%	13.76%
lex1	**56.04%**	7.03%	12.49%
lex2	47.95%	5.66%	10.12%
lex3	34.05%	8.73%	13.90%
cgp1	25.70%	24.55%	25.11%
cgp2	36.35%	18.28%	24.33%
cgp3	33.13%	3.79%	6.80%
pos1	42.42%	0.97%	1.90%
pos2	42.66%	4.21%	7.66%
pos3	38.75%	2.14%	4.06%

Table 1: Results of *one vs. all* Authorship Verification experiments using MBL

Table 1 shows the results for the positive class of *one vs. all* authorship verification using memory-based learning. We see that memory-based learning on the authorship verification task is able to correctly classify up to 56% of the positive class which is highly underrepresented in both training and test data. Despite the very skewed class distributions, memory-based learning scores reasonably well on this approximation of authorship verification with limited data. The most important lesson is that in a realistic set-up of the task of authorship verification, the accuracy to be expected is much lower than what in general can be found in the published literature.

5 Related Research

As mentioned earlier, most research in authorship attribution starts from unrealistic assumptions about numbers of authors and amount of training data available. We list here the exceptions to this general rule. These studies partially agree with our own results. Argamon et al. (2003) report on results in authorship attribution on twenty authors in a corpus of Usenet newsgroups on a variety of topics. Depending on the topic, results vary from 25% (books, computer theory) to 45% accuracy (computer language) for the 20-author task. Linguistic profiling, a technique presented by Van Halteren (2005), takes large numbers of linguistic features to compare separate authors to average profiles. In a set of eight authors, a linguistic profiling system correctly classifies 97% of the test documents. Madigan et al. (2005) use a collection of data released by Reuters consisting of 114 authors, each represented by a minimum of 200 texts. Results of Bayesian multinomial logistic regression on this corpus show error rates between 97% and 20%, depending on the type of features applied. This is only partially comparable to the authorship attribution results on 145 authors presented in this paper because of the large amount of data in the Madigan et al. (2005) study, while our system works on limited data. In a study of weblog corpora, Koppel et al. (2006) show that authorship attribution with thousands of candidate authors is reasonably reliable, since the system gave an answer in 31.3% of the cases, while the answer is correct in almost 90% of the cases. Whereas these cases show similar results as ours, we believe this study is the first to study the effect of training set size and number of authors involved systematically.

When applied to author verification on eight authors, the linguistic profiling system (Van Halteren, 2005) has a False Reject Rate (FRR) of 0% and a False Accept Rate (FAR) of 8.1%. Argamon et al. (2003) also report on *one vs. all* learning in a set of twenty authors. Results vary from 19% (books, computer theory) to 43% (computer language) accuracy, depending on the topics. Madigan et al. (2005) also did authorship verification experiments on their corpus of 114 authors we described above. They vary the number of target, decoy, and test authors to find that the ideal split is 10-50-54, which produces an error rate of 24%. Koppel et al. (2007) also report on results in *one vs. all* experiments. Using a corpus of 21 books by 10 authors in different genres (including essays, plays, and novels), their system scores a precision of 22.30% and recall of 95%. Our system performs better in precision and worse in recall. Their corpus nevertheless consists of 21 books (each represented by more than forty 500-word chunks) by 10 authors, which makes the task considerably less difficult.

6 Conclusions and Further Research

A lot of the research in authorship attribution is performed on a small set of authors and unrealistic sizes of data, which is an artificial situation. Most of these studies not only overestimate the performance of their system, but also the importance of linguistic features in experiments discriminating between only two or a small number of authors. In this paper, we have shown the effect of many authors and limited data in authorship attribution and verification. When systematically increasing the number of authors in authorship attribution, we see that performance drops significantly. Similar types of features work well for different amounts of authors in our corpus, but generalizations about individual features are not useful.

Memory-based learning shows robustness when dealing with limited data, which is essential in e.g. forensics. Results from experiments in authorship attribution on 145 authors indicate that in almost 50% of the cases, a text from one of the 145 authors is classified correctly. Using combinations of good working lexical and syntactic features leads to significant improvements. The authorship verification task is a much more difficult task, which, in our approximation of it, leads to a correct classification in 56% of the test cases. It is clear that studies reporting over 95% accuracy on a 2-author study are overestimating their performance and the importance of the features selected.

Further research with the 145-author corpus will involve a study of handling with imbalanced data and experimenting with other machine learning algorithms for authorship attribution and verification and a more systematic study of the behavior of different types of learning methods (including feature selection and other optimization issues) on this problem.

Acknowledgements

This study has been carried out in the framework of the 'Computational Techniques for Stylometry for Dutch' project, funded by the National Fund for Scientific Research (FWO) in Belgium. We would like to thank the reviewers for their comments which helped improve this paper.

References

Argamon, Shlomo, Marin Saric, and Sterling S. Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: First results. In *Proceedings of the 2003 Association for Computing Machinery Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pages 475–480.

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sushant Dawhle, Sobhan R. Hota, Navendu Carg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science and Technology*, 58(6):802–822.

Baayen, Harald R., Hans Van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131.

Briggs Myers, Isabel and Peter B. Myers. 1980. *Gifts differing: Understanding personality type*. Mountain View, CA: Davies-Black Publishing.

Burrows, John. 2007. All the way through: Testing for authorship in different frequency data. *Literary and Linguistic Computing*, 22(1):27–47.

Coyotl-Morales, Rosa M., Luis Villaseñor Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. Authorship attribution using word sequences. In *Proceedings of the Iberoamerican Congress on Pattern Recognition (CIARP)*, pages 844–853.

Daelemans, Walter and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2007. TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide. Technical Report ILK Research Group Technical Report Series no. 07-07, ILK Research Group, University of Tilburg.

Diederich, Joachim, Jörg. Kindermann, Edda Leopold, and Gerhard Paass. 2000. Authorship attribution with Support Vector Machines. *Applied Intelligence*, 19(1-2):109–123.

Feiguina, Ol'ga and Graeme Hirst. 2007. Authorship attribution for small texts: literary and forensic experiments. In *Proceedings of the 30th International Conference of the Special Interest Group on Information Retrieval: Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (SIGIR)*.

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

Hirst, Graeme and Ol'ga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417.

Holmes, D. 1994. Authorship Attribution. *Computers and the Humanities*, 28(2):87–106.

Koppel, Moshe, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th International Conference of the Special Interest Group on Information Retrieval (SIGIR)*, pages 659–660.

Koppel, Moshe, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, 8:1261–1276.

Le, Zhang. 2006. Maximum Entropy Modeling Toolkit for Python and C++. Version 20061005.

Luyckx, Kim and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Madigan, David, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*.

Miranda García, Antonio and Javier Calle Martín. 2007. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1):49–66.

Mosteller, F. and D. Wallace. 1964. Inference and disputed authorship: the Federalist. *Series in Behavioral Science: Quantitative Methods Edition*.

Platt, John, 1998. *Advances in Kernel Methods - Support Vector Learning*, chapter Fast training of Support Vector Machines using Sequential Minimal Optimization, pages 185–208.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *Association for Computing Machinery (ACM) Computing Surveys*, 34(1):1–47.

Stamatatos, Efstathios. 2007. Author identification using imbalanced and limited training texts. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA)*, pages 237–241.

Van Halteren, Hans. 2005. Linguistic profiling for author recognition and verification. In *Proceedings of the 2005 Meeting of the Association for Computational Linguistics (ACL)*.

Witten, Ian and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools with Java Implementations*. San Fransisco: Morgan Kaufmann.