

Author Profiling for English Emails

**Dominique Estival, Tanja Gaustad,
Son Bao Pham, Will Radford**
Appen Pty Ltd
Chatswood NSW 2067, Australia
{destival,tgaustad,sbpham,wradford}@appen.com.au

Ben Hutchinson*
Google Australia
Sydney NSW 2000, Australia
benhutch@google.com

Abstract

This paper reports on some aspects of a project aimed at automating the analysis of texts for the purpose of author profiling and identification. The complete analysis provides probabilities for the author's basic demographic traits (gender, age, geographic origin, level of education and native language) as well as for five psychometric traits. We describe the email data which was collected for the project, the ways this data is processed and analysed, and the experimental setup used for classification with the Text Attribution Tool (TAT) before presenting our results for the demographic and psychometric traits using English email. Results are very promising for all ten traits examined.

1 Introduction

Automatically predicting the identity of authors from their texts has a number of potential applications. For example, if a text poses any type of threat, then identifying the source of the threat is the first step in countering it. In this context, author profiling forensics can be helpful to at least narrow the list of potential authors (Corney et al., 2002; Argamon et al., 2005; Abbasi and Chen, 2005). Another area where author identification and profiling can provide valuable information is in deriving marketing intelligence from the acquired profiles (Glance et al., 2005) and the rapidly growing field of sentiment analysis and classification (Oberlander and Nowson, 2006).

In this paper, we present some aspects of a research project aimed at automating the analysis of text in various forms of data (email in the

first instance), for the purposes of author profiling and identification. The complete analysis provides probabilities for the author's basic demographic traits such as gender, age, geographic origin, level of education and native language, as well as for five psychometric traits. The prototype software¹ also provides a probability of a match with other texts, both from known and unknown authors. This paper describes experiments on profiling authors of emails in English using various machine learning (ML) algorithms and feature sets.

We will first give an overview of previous work done in authorship and text attribution and our specific take on the problem (Section 2). In Section 3, we introduce the specifics of our data set, including information on the data collection process, normalisation and validation as well as the in-depth analysis performed on the data. After a description of the experimental setup in Section 4, we report on and discuss our results for five demographic and five psychometric traits in Section 5. Section 6 concludes the paper with some pointers to future experiments.

2 Author attribution and author profiling

Authorship attribution is the task of deciding for a given text which author (usually from a predefined set of authors) has written it. Classic examples include authorship attribution studies on the Bible (Friedmann, 1997), Shakespeare's works (Ledger and Merriam, 1994) or the Federalist Papers (Mosteller and Wallace, 1964). For a long time, the main applications have been restricted to literary texts. Recently, authorship attribution has gained new life in the fight against cyber crime and in a more general search for reliable identification techniques (Abbasi and Chen, 2005; de Vel et al.,

The work presented in this paper was carried out while this author was working at Appen Pty Ltd.

¹The working prototype of TAT has now been completed and delivered, but will not be discussed here in more detail.

2001, 2002; Zheng et al., 2003).

The task of authorship attribution has traditionally been carried out on data from small sets of authors. For larger data sets, involving more authors, the challenge of identifying individual authors is more difficult. In such cases, predicting characteristics, or traits, of authors can be a good alternative and provide clues as to the author's identity.

Author profiling is the task of predicting one or more such author traits and an author profile consists of the resulting set of one or more predicted traits. Importantly, and contrary to author attribution, the author profiling task is possible even when documents by the author are not in the training data. Also in contrast to author attribution, greater accuracy can be expected when the training data contains texts from more authors, because the models of each trait are then expected to be more robust.

Most research into author profiling focuses on the prediction of a small number of traits, e.g. gender (Corney et al., 2002), gender and age (Koppel et al., 2006), neuroticism and extraversion² (Argamon et al., 2005), neuroticism, agreeableness, extraversion and conscientiousness (Oberlander and Nowson, 2006), neuroticism, agreeableness, extraversion, conscientiousness, and openness (Mairesse and Walker, 2006). In contrast, the present study is the largest that we know of in terms of the number of traits predicted, ten in total (see Section 4).

Corney et al. (2002) describe an experiment in predicting gender in email with a machine learner, namely SVM. The majority of the features the authors use are similar to the ones employed in our system with the exception of the gender-preferential linguistic features. Overall, this approach satisfactorily discriminates between male and female authors. The main finding of Corney et al. is that function words provide the most important clues for differentiating gender. This study is most comparable to the work presented in this paper.

Argamon et al. (2005) try to distinguish high neuroticism from low neuroticism and extraversion from introversion in informal texts. The approach uses four sets of features (lexical features, conjunctive phrases, modality, appraisal) and an

²Both "extraversion" and "extroversion" can be used to describe this particular dimension of human personality. In the domain of personality models, however, "extraversion" is the accepted term.

SMO machine learner for classification. Even though the authors define the task as a binary classification task distinguishing between the top third and the bottom third scores for the two psychometric traits, the results are inconclusive. The authors conclude that most probably the features chosen are not adequate for the task.

Koppel et al. (2006) predict gender and age in blog data. Due to the massive number of authors investigated (18,000), a regular classification approach is not feasible. The authors opted for an Information Retrieval technique using various term frequency-inverse document frequency weights in combination with a cosine measure for similarity. In approximately 70% of the attempted predictions, this method is not able to pick an author for a given blog. Prediction accuracy for the blogs that have been assigned an author reaches 88.2%. Because of the very different approach, the use of different features and the fact that in their study the two traits are evaluated in combination, it is impossible to compare this approach to ours.

Another blog study which predicts four psychometric traits (neuroticism, agreeableness, extraversion and conscientiousness) using machine learners (Naive Bayes and SVM) is presented in (Oberlander and Nowson, 2006). The blog corpus consists of 71 authors (notably smaller than the data used in (Koppel et al., 2006)) and the authors use word bi- and trigrams as features. They report on many different setups based on binary or n-ary classification and different levels of restrictions on feature selection. Of the tasks described, task 6 using a 3-way split and automatic feature selection is most similar to the setup used in our experiments. Oberlander and Nowson achieve promising results for all the traits examined.

Mairesse and Walker (2006) try to predict five psychometric traits in order to build reliable personal profiles for dialogue management. They use various machine learners on two corpora, one written and one spoken, with features based on speech acts (command, prompt, question, assertion), the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2001) (e.g. ratio of pronouns, positive emotion words), the MRC Psycholinguistic database (Coltheart, 1981) (frequency of word use, familiarity, age of acquisition) and—for the spoken corpus—prosody (voice pitch and intensity, speech rate). The features used in their study are very different from those used for the exper-

iments described in this paper. Unfortunately, Mairesse and Walker (2006) do not include baselines in their paper and we can therefore not compare our performance to theirs. The main conclusions are that observed personality is easier to model than self-reports and that spoken language is easier to model than written language.

3 Data

The data we are using for the experiments described here are email messages which have been collected specifically for the purpose of the TAT project. Because this constitutes a new data set, we include a discussion of the data collection methodology, the normalisation and validation of the email messages before their inclusion in the corpus, as well as the analysis and more in-depth pre-processing of the actual text.

3.1 Data Collection and Validation

We collected emails in several varieties of English, including native and non-native speakers of English coming from different geographical areas. Table 1 gives an overview of the different types of emails written in English, with statistics for the number of authors, number of emails and total number of words in the corpus.

Respondents were contacted via a recruitment process which included notification of privacy and assurance that their identity would be protected. Respondents agreed to fill out a web questionnaire and to donate ten email messages. The questionnaire elicited information about both demographic and psychometric traits. Demographic traits cover basic demographic information about the author: age, gender, native language, level of education, and main country of residence. For the psychometric traits, all the English collections (with the exception of the Egyptian one) use a short version of the International Personality Item Pool questionnaire, consisting of 41 questions (Buchanan et al., 2005).

After completing the questionnaire, respondents forwarded previously sent emails, e.g. from their email client “SentBox”, to the data collection email address. The raw email documents were stored on a dedicated mail server and from then on all further operations took place on copies. The email messages were first checked manually to filter out erroneous content such as foreign language emails or forwarded chain letters and to ensure

consistency and accuracy of the documents in the corpus. Email messages containing less than 5 lines of text (at a fixed line length of 80 characters) were excluded from our data set.

Data from authors who did not meet a set of satisfaction criteria was also removed. These criteria were: 1) a valid questionnaire received for a given author; 2) at least 5 valid email messages for the author; and 3) a total word count for that author’s valid email messages of at least 1000 words. An additional requirement was that the emails be from different domains, such as personal or business emails. In the end, 1033 respondents were validated with a combined total of 9836 email messages. This corresponds to 57% of the total number of emails collected and constitutes the data set used for the experiments reported in this paper.

3.2 Normalisation

Forwarded emails come in a variety of formats and pre-processing was required to normalise those. This includes resolving encoding issues, identifying the actual content part of the email message and removing artefacts introduced through the data collection methodology.

The process takes into account languages other than English and writing systems other than Latin script. Ultimately, all the email documents are stored in UTF-8 encoding. However, the original documents come in a variety of character set encodings, e.g. US-ASCII, UTF-8, ISO-8859-1. Moreover, some documents have inconsistent encoding declarations, e.g. a UTF-8 document may have a header claiming it is ISO-8859-1. Such cases are handled using a set of heuristics to guess at the correct encoding.

Parts of the MIME message containing the body of the email are distinguished from headers, attachments and forwarded material sent as embedded MIME messages. Non-text/non-HTML payload contents, or anything with a file name, are also stripped from the version of the documents to be processed.

3.3 Analysis

The aim of the analysis phase is to produce annotations which are later used during the feature extraction phase. The analysis stage consists of three modules: document parsing³, text processing and

³In order to avoid confusion, it is important to stress that we are using “parsing” in the computer science interpretation of the term and not as in “syntactic parsing”.

Collection	Native lang.	# authors	# emails	# words total	# words by author
United States	English	415	4,533	2,405,792	1,886,389
United Kingdom	English	23	273	178,400	137,238
Australia/New Zealand	English	133	1,387	513,065	437,454
United States	Spanish	174	1,823	519,504	461,767
Egypt	Arabic	288	1,820	451,903	444,325
Total		1,033	9,836	4,068,664	3,367,173

Table 1: Overview of the collected English email data.

linguistic analysis, each resulting in a different set of annotations.

There are two ways annotations can be created: automatically, by the analysis modules described below, and manually by human annotators according to pre-specified guidelines using the Callisto tool (Mitre, 2006). Our emphasis is on automatically created features, mainly using the manual annotation for validation purposes.

3.3.1 Document parsing

The purpose of the document parsing stage is to classify the body text in an email document into five categories:

1. Author text: text that was written by the author and that is not contained in an embedded reply chain of email messages;
2. Signature: email signature text, which typically includes contact information, professional details, and/or quotations;
3. Advertisement: advertisements automatically appended by the author’s email client, such as Yahoo and Hotmail;
4. Quoted text: extended quotations, e.g. song lyrics, poems, newspaper articles;
5. Reply lines: text that was written in a previous email message that the author is either forwarding or replying to, including text by other writers, text in previous emails by the author of the current email, with their email signatures, advertisements and quotations.

The input is an email document and the output is the same email document, with the lines of the body text categorised into one of those five classes. Consecutive lines of the same class are covered by an annotation of the corresponding type. These

annotations are used to calculate structural features.

The document parsing stage is crucial, as it is the linguistic features of the “author text” which provide the most clues for author attribution. Table 1 gives the number of words written by the author in the body of the email messages and the total number of words contained in the emails.

We experimented with an existing tool, Jangada (Carvalho and Cohen, 2004), which identifies signature blocks and replies, but unfortunately it did not perform very well on our data and we identified some shortcomings. The main issues were that we could not configure Jangada to distinguish the more detailed categories enumerated above nor include additional features in its (rather simple) statistical model of document structure. Also, in spite of what seems to be suggested in (Carvalho and Cohen, 2004), Jangada makes systematic errors and does not identify forwarded message text as reply lines.

The poor performance of Jangada led us to develop our own document parser. This document parser builds a statistical model of document structure by extracting features from each line of a document and using them to train a statistical classifier. Conditional Random Fields (Lafferty et al., 2001) have been shown to work well at labeling sequential data as they can effectively combine contextual information and line-specific features and have therefore been employed as a statistical model in our system.

To compare the performance of Jangada with that of our document parser, we used ten-fold cross validation on all of the English data. For each train-test partition, our document parser was trained on the training part and its performance was compared with Jangada’s on the testing part. As Jangada can only identify *Author text*, *Reply* and *Signature* lines, we cannot compare the per-

formance on the task of recognising all five categories. For the task of identifying those three categories, our document parser achieved an accuracy of 88.16% while Jangada performed at 64.22%. When focusing on the task of identifying only author lines, our document parser reached an F-score of 90.76% compared to 74.64% from Jangada.

3.3.2 Text processing

Text processing consists of two stages: segmentation and punctuation analysis. First, the text in the email is split into paragraphs and paragraphs are split into sentences and tokens. The latter is currently performed with third party tools (Cunningham et al., 2002) which generate *paragraph*, *sentence* and *token* annotations respectively. The use of sentence punctuation marks and other special characters is then analysed, including but not limited to, special markers, e.g. two hyphens “- -” which often indicate that an email signature follows; quotation marks, which sometimes signal the presence of a quotation; and emoticons, such as “:-)” or “:o)”. This information is stored as attributes of *token* annotations which are used for calculating character level features.

3.3.3 Linguistic analysis

The aim of the linguistic analysis stage is to produce more linguistically informed annotations, such as Part-Of-Speech (POS) tags. Linguistic processing deals with the linguistic, usually language-dependent, aspects of texts and it requires linguistic resources such as word lists.

To identify key phrases, we have developed a Named Entity recognizer using gazetteers and grammars, which identifies people, locations, organisations, dates etc. We implemented our own NE recognizer because most available systems were developed on news corpora, i.e. a very different domain from ours. All the heuristics in our recognizer are based on email data. Additional lexicons were developed manually to identify set phrases, for instance farewells and greetings.

4 Experimental setup

Many problems in NLP have lent themselves to solutions using statistical language processing techniques. Author profiling can be considered a type of document classification task, where the classes correspond to traits of the authors. These traits are arranged along various dimensions, with

different options for each dimension being mutually exclusive. For example male and female are the possibilities for the gender dimension.⁴ For each dimension, the email and questionnaire data are used to construct classifiers, using a range of ML techniques.

Each document constitutes a single data instance for the purposes of the experiments. For each experiment, ten-fold cross-validation was used, so the results reported in Section 5 are on the entire data set. We also used ten-fold cross-validation during training, for feature selection and model parameter tuning. This means that for each train-test partition in cross-validation, we tuned the parameters on the training part using ten-fold cross-validation. Once the best combination of ML classifiers, parameters and feature selection has been determined, that model is used to classify the test data to evaluate the performance of the chosen model.

4.1 Traits and classes

We distinguish five demographic and five psychometric traits in the experiments presented in this paper, namely *age*, *gender*, *native language*, *level of education* and main *country* of residence for the demographic traits, and *agreeableness*, *conscientiousness*, *extraversion*, *neuroticism* and *openness* for the psychometric traits (Matthews et al., 2003). This information is extracted from the questionnaire filled out by the respondents.

For the traits taking numerical values, subjects were split into three classes based on the first and third quartiles. Table 2 summarizes the data distribution for each trait across these classes.

4.2 Features

For each author, a feature vector is calculated. Typically, a feature is a descriptive statistic calculated from both the raw text and the annotations. For example, a feature might express the relative frequency of two different annotation types (e.g. number of words/number of sentences), or the presence or absence of an annotation type (e.g. signature).

For the English data, 689 features were calculated. These were divided into three main groups,

⁴It has been pointed out that a male author could also be writing from a female perspective and vice versa. In our corpus, however, these cases do not occur and therefore gender is treated as a binary variable.

Demographics				
Age:	Gender:	Native language	Level of education:	Country:
<25 (423)	Male (483)	English (571)	No tertiary edu. (498)	USA (528)
25 to 35 (350)	Female (550)	Arabic (288)	Some tert. edu. (535)	Egypt (288)
>35 (260)		Spanish (174)		AUS/NZ (133)
				Other (84)
Psychometrics				
Agreeableness:	Extraversion:	Neuroticism:	Conscientiousness:	Openness:
<5	<1	<-7	<3	<2
5 to 8	1 to 7	-7 to -2	3 to 9	2 to 7
>8	>7	>-2	>9	>7

Table 2: Traits and Classes, with frequencies in parentheses where applicable.

namely character-level, lexical, and structural features. In turn, we subdivided these main groups into feature groups which subsume all features of a particular type. An overview of the feature groups is shown in Table 3. The main purpose of these groupings was to make more informed choices during the feature selection stage and to facilitate experimentation with various combinations of feature groups. For instance “char+lexical” denotes the combination of all character-level and all lexical-level features, whereas “all-html” expresses that all features except the html group were used during classification.

Character-level features cover features such as the frequency of punctuation characters or word length. New types of features for this group include case-based features relating to occurrences of CamelCase and slow Shift release, e.g. “Hello”.

The lexical feature group comprises for instance function words and POS. A special addition to this group are features based on words that are highly correlated with one of the values for a particular author trait. For example, younger people use more first person pronouns whereas older people use more third person pronouns, and non-US people are more likely to use the abbreviations “u” and “ur” (for “you” and “your”).

At the structural level, we included features such as paragraph breaks and the presence or absence of certain HTML tags. The annotations created by our document parser also produced features which identify interleaved author and reply text, as opposed to a consecutive text structure.

4.3 Classification algorithms and feature selection

The aim of the classifier is to match feature vectors from the document with author traits. Ordered pairs of feature vectors and author traits are used to train and tune machine learning classifiers. Formally, classifiers are functions which map feature vectors to author traits and there will be classifiers for each author trait such as gender, age, etc.

We apply various machine learning algorithms as classifiers, using the WEKA toolkit (Witten and Frank, 2005) to find the best classifier for each trait. During training, classifiers are created by the selection of sets of features for each author trait, and classifier parameters are tuned through cross-validation. To evaluate and test the classifiers, new documents are given as input and existing classifiers are selected to predict author traits.

The machine learning algorithms we tried include decision trees (J48 (Quinlan, 1993), RandomForest (Breiman, 2001)), lazy learners (IBk (Aha et al., 1991)), rule-based learners (JRip (Cohen, 1995)), Support Vector Machines (SMO (Keerthi et al., 2001), LibSVM (Chang and Lin, 2001)), as well as ensemble/meta-learners (Bagging (Breiman, 1996), AdaBoostM1 (Freund and Schapire, 1996)). These algorithms were used in combination with feature selection methods based on either a feature subset evaluator together with a search method (consistency subset evaluator with a best-first search) or a single attribute evaluator with various numbers of attributes selected (χ^2 , GainRatio, and InformationGain) (see chapter 10.8 in (Witten and Frank, 2005) for details).

Main group	Feature group	Description
character		Features at character level
	case	Features using <i>case</i> attributes of characters
	wordLength	Features invoking word length
lexical		Features at lexical level
	functionWord	Features invoking function words
	correlate	Features using words that are highly correlated with a trait class
	namedEntities	Features using named entities
	POS	Features based on POS
structural		Features at structural level
	docCategory	Features specifying the category of an email (e.g. personal)
	html	Features pertaining to the html rendering of the email

Table 3: Feature groups.

5 Results and discussion

The results shown here were computed on the English email data set described in Section 3 using the different classifiers and general setup introduced in Section 4. Table 4 shows the results on all ten traits (demographic and psychometric). It also includes the respective baseline associated with each separate classification task, calculated on our data set of 9,836 emails. Furthermore, we state which settings (ML algorithm, feature selection, and feature set) were used to achieve the results reported. Education and gender are both binary classification tasks, whereas age and native language have three classes and country of residence four classes. All the psychometric traits are divided into three classes (see Section 4.1 and Table 2 for details on the exact split).

These results show that for the demographic and the psychometric author profile, classification is significantly⁵ improved over the baseline for all ten traits. This demonstrates that the approach we took of combining ML algorithms, together with our particular feature set, is successful for binary as well as n-ary classifications on very diverse classification tasks.

No clear picture emerges as to which ML algorithms perform best over all the classification tasks. Our results seem to indicate that SMO works well for three out of five demographic traits, and IBk shows good performance for three out of five psychometric traits.

With regard to feature selection, our expectation was that for a Support Vector Machines algorithm, such as SMO, feature selection should not make

a huge difference whereas an algorithm based on decision trees, such as RandomForest, would be more sensitive to feature selection in general. This expectation was, however, not confirmed during our experiments.

Looking at the features used for the predictions, taking all available features (with or without the exclusion of a certain feature group) works best for all demographic traits. On the other hand, all psychometric traits except openness rely on the combination of character-based and structural features; openness relies on structural features alone.

It is especially interesting to see that for the prediction of level of education, explicitly excluding function words from the features used leads to the best performance. Previous publications have found that function words are surprisingly effective on their own for authorship attribution (Argamon and Levitan, 2005; de Vel et al., 2002) or for certain traits, such as gender (Corney et al., 2002). However, in our data set, function words do not seem to work that well for predicting level of education

Comparing our findings to previously published results shows the following. For gender, our system performs very close to that of Corney et al. (2002): our relative error is 68.3 whereas their results yield 69.2 relative error. The results on psychometric traits reported in (Oberlander and Nowson, 2006) are very high and our system does not achieve the same performance, but one has to bear in mind that our corpus is 14 times larger than theirs and consists of email data rather than blogs. It is worth noting that on extraversion, our relative error slightly exceeds theirs.

Future research will need to investigate deeper

⁵All results are significant at $p=0.01$ using a χ^2 test.

Trait	ML algorithm	Feature Sel.	Best Features	Results	Baseline
Age:	SMO	–	all	56.46	39.43
Gender:	SMO	–	all	69.26	54.48
Language:	RandomForest	InfoGain	all-correlate	84.22	62.90
Education:	Bagging	–	all-functionWord	79.92	58.78
Country:	SMO	–	all	81.13	57.29
Agreeableness:	IBk	–	char+structural	53.16	40.51
Conscientiousness:	IBk	–	char+structural	54.35	43.72
Extraversion:	LibSVM	–	char+structural	56.73	45.17
Neuroticism:	IBk	–	char+structural	54.29	42.34
Openness:	RandomForest	–	structural	55.32	47.28

Table 4: Results for all demographic and psychometric traits on English email data.

features, such as syntactic information or writing style, which might help to classify the author traits more accurately. It would also be interesting to identify more specialised feature sets for each author trait. As mentioned above, a first step in this direction was made using words that correlate with a particular instantiation of a trait, but more research exploring features tailored to a given trait is necessary.

6 Conclusion and future work

We have presented some results of experiments to automatically predict author traits from email messages. This work is of interest for a number of potential applications, from threat identification to marketing intelligence. The results presented in this paper were conducted on the English subset (9836 emails) of the email data we have collected.

The experiments reported here were aimed at discovering how well a range of ML algorithms perform on our data set for five demographic and five psychometric author traits. Our results show that the chosen approach works well for author profiling and that using different classifiers in combination with a subset of available features can be beneficial for predicting single traits.

The next steps are to extend the text processing modules to other languages and to conduct a more thorough error analysis for our results. In particular, a more qualitative analysis per instance instead of overall performance will definitely provide more insight. In addition, it would be interesting to extend our approach to the task of author identification.

The email corpus presented in this paper has been collected in a principled manner and will

hopefully become a valuable resource for the community.⁶

Acknowledgements

This research was carried out within the framework of a US Government BAA grant (IS-QD-2467). We would like to thank Judith Bishop for her efforts during data collection, as well as the anonymous ACL and PACLING reviewers for their valuable comments and insights on a previous version of this paper.

References

- Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to Arabic web content. In Paul B. Kantor et al., editor, *Intelligence and Security Informatics, Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2005)*, pages 183–197. Springer.
- David Aha, Dennis Kibler, and Mark Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.
- Shlomo Argamon, Sushant Dhawle, Mosche Koppel, and James W. Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*. St. Louis.
- Shlomo Argamon and Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*. Victoria, BC.

⁶Please contact one of the authors at Appen for more information on the email corpus.

- Leo Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2):123–140.
- Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.
- Tom Buchanan, John A. Johnson, and Lewis R. Goldberg. 2005. Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment*, 21:115–127.
- Vitor Carvalho and William Cohen. 2004. Learning to extract signature and reply lines from email. In *Proceedings of the Conference on Email and Anti-Spam (CEAS-2004)*. Mountain View.
- Chih-Chung Chang and Chih-Jen Lin. 2001. *LIB-SVM: a library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- William Cohen. 1995. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123.
- Max Coltheart. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.
- Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC 2002)*, pages 282–292. Las Vegas.
- Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175.
- Olivier de Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining email content for author identification forensics. *SIGMOD Record*, 30(4):55–64.
- Olivier de Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2002. E-mail authorship attribution for computer forensics. In Daniel Barbara and Sushil Jajodia, editors, *Data Mining for Security Applications*. Kluwer Academic Publishers.
- Yoav Freund and Robert Schapire. 1996. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156. San Francisco.
- Richard Friedmann. 1997. *Who wrote the Bible?* Harper, San Francisco.
- Natalie Glance, Matthew Hurst, Kamal Nigam, Mathew Siegler, Robert Stockton, and Takashi Tomokiyo. 2005. Deriving marketing intelligence from online discussion. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428. Chicago.
- Sathiya Keerthi, Shirish Shevade, C. Bhattacharyya, and K. Murthy. 2001. Improvements to Platt’s SMO algorithm for SVM classifier design. *Neural Computation*, 13(3):637–649.
- Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of SIGIR*, pages 659–660.
- John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco.
- Gerrard Ledger and Thomas Merriam. 1994. Shakespeare, Fletcher, and The Two Noble Kinsmen. *Literary and Linguistic Computing*, 9(3):235–248.
- François Mairesse and Marilyn Walker. 2006. Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*, pages 543–548. Vancouver.
- Gerald Matthews, Ian Deary, and Martha Whiteman. 2003. *Personality Traits*. Cambridge University Press, Cambridge, second edition.
- Mitre. 2006. Callisto. <http://callisto.mitre.org/>. Version 1.4.0.
- F. Mosteller and D.L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Series in behavioral science: Quantitative methods edition. Addison-Wesley.
- Jon Oberlander and Scott Nowson. 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634. Sydney.

- James Pennebaker, Martha Francis, and Roger Booth. 2001. LIWC: Linguistic inquiry and word count. <http://www.liwc.net/>.
- Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo.
- Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edition.
- Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. 2003. Authorship analysis in cybercrime investigation. In Hsinchun Chen et al., editor, *Proceedings of the first NSF/NIJ Intelligence and Security Informatics Symposium*, pages 59–73. Springer.