Proceedings of the Australasian Language Technology Workshop 2007, pages 21-30

TAT: an author profiling tool with application to Arabic 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
Sydney, Australia
benhutch@google.com

Abstract

This paper reports on the application of the Text Attribution Tool (TAT) to profiling the authors of Arabic emails. The TAT system has been developed for the purpose of language-independent author profiling and has now been trained on two email corpora, English and Arabic. We describe the overall TAT system and the Machine Learning experiments resulting in classifiers for the different author traits. Predictions for demographic and psychometric author traits show improvements over the baseline for some of the author traits with both the English and the Arabic data. Arabic presents particular challenges for NLP and this paper describes more specifically the text processing components developed to handle Arabic emails.

1 Introduction

The goal of the TAT project is to develop a language-independent Text Attribution Tool (TAT) which can provide information on authors for a variety of document types and a range of languages. In the first implementation, the TAT has been developed for profiling the authors of email messages in English and Arabic, with other languages to be added in the future. (Estival et al., 2007) describes the Machine Learning experiments and the results obtained for the English email data. In this paper, we will focus on the results obtained for Arabic emails and will describe the aspects of the TAT which are particular to the processing of our Arabic email data.

We first introduce the two tasks of author attribution and author profiling in Section 2. In Section 3, we describe the data set on which our tool was trained and tested. Section 4 contains an overall description of the TAT system, followed by a more specific description of the processing steps required for Arabic emails in Section 5. After presenting the general experimental setup in Section 6, we report on the results achieved for several demographic and psychometric traits in Section 7 and propose our general conclusions in Section 8.

2 Author attribution and author profiling

The ability to recognise the identity or certain characteristics of authors automatically from texts has a number of potential applications. Author identification and author profiling can provide valuable information for marketing intelligence (Glance et al., 2005), while the rapidly growing field of sentiment analysis and classification (Oberlander and Nowson, 2006) is another application where profiling can make a contribution. Also, author profiling forensics may be helpful in narrowing the choice of potential authors when identifying the source of a threat (Corney et al., 2002; Argamon et al., 2005; Abbasi and Chen, 2005a).

Author attribution is the task of deciding for a given text which author, usually from a predefined set of authors, has written it. Historically, author identification has its roots in literature, with studies of the Bible (Friedmann, 1997), Shakespeare (Ledger and Merriam, 1994) or the Federalist Pa-

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

pers (Mosteller and Wallace, 1964). Recently, author identification has also been applied to more informal texts, such as emails (de Vel, 2000; de Vel et al., 2001; de Vel et al., 2002), newsgroup messages (Zheng et al., 2003; Zheng et al., 2006) or blog entries (Koppel et al., 2005; Oberlander and Nowson, 2006).

Author profiling is the task of predicting one or more traits for the author of a text and the author profile consists of the set of traits predicted for that author. A major difference between author profiling and author attribution is that it is possible to predict author traits even when the training data does not contain any document by the actual author. Another difference is that greater accuracy can be expected for author profiling when the training data is made up of more authors, because the models for each trait are expected to be more robust. The accuracy of author identification, on the other hand, can be expected to decrease if the number of potential authors is increased.

In (Estival et al., 2007), we presented our work on author profiling for English emails and discussed the literature on previous research in that area for English texts. For Arabic texts, very few studies have been published in the area of author attribution; the only work we know of can be found in (Abbasi and Chen, 2005b) and (Abbasi and Chen, 2005a). We are not aware of any work on author profiling for Arabic.

For author attribution in Arabic, Abbasi and Chen (2005a) apply two different machine learners to a dataset of Arabic forum messages by 20 authors (with 20 messsages per author). The feature set comprises lexical, syntactic, structural and content-specific features, including a number of features specifically tailored to Arabic. These relate mostly to inflection (counting roots rather than inflected words), word length (adjusting the range for the Arabic word length distribution to reflect the fact that Arabic words tend to be shorter than English ones) and elongation dashes (excluding them from word length measurements, but tracking their usage). The main conclusion reached is that using an SVM classifier with all the features achieves the best accuracy on their data set, but that the overall performance is lower than for English.

Abbasi and Chen (2005b) use the same approach, and in addition include an in-depth comparison between the feature sets and results for English and Arabic on forum messages.

3 Data

The data collected for this project and used for training the TAT consists of two sets of emails from 1033 English speakers and from 1030 Arabic speakers. The English data set contains emails written by both native and non-native speakers of English: native speakers of US English, native speakers of Australian/NZ English, native speakers of Spanish from the US and native speakers of Egyptian Arabic from Egypt. The Arabic data set contains emails written by native speakers of Egyptian Arabic, as described in more detail below. In the rest of this section, we focus on the data collection and validation processes for Arabic emails.

3.1 Data Collection

For the Arabic email data set, the collection was conducted in Egypt and all the writers were native speakers of Egyptian Arabic. Compared with the English email data set, a special feature is the encoding of the input for Arabic email. The widespread use of the Internet and even more of text messages via mobile phones without the possibility of Arabic script input has led to the use of the Latin alphabet and the development of some transliteration conventions in the Arabic speaking world. Even though Arabic keyboards are now more common, people still sometimes write email using a Latin keyboard. However, there are no strict rules for the conversion of Arabic script into Latin characters and the way of writing emails with a Latin keyboard varies greatly according to dialect or country, and even across individuals.

Table 1 gives an overview of the number of authors, number of emails and number of words for both the Arabic and English data set. We also include the proportion of emails in Arabic script and in Latin script for the Arabic data.

The data collection process for the Arabic data differed slightly from the process described in (Estival et al., 2007) for the collection of the English data, in that the respondents had to come to a central collection location. Nevertheless, the recruitment pro-

Collection	# authors	# emails	# words
Arabic	1,030	8,028	2,153,333
Arabic script		7,267	
Latin script		761	
English	1,033	9,836	3,367,173

Table 1: Overview of the collected English and Arabic email data.

cess also included notification of privacy and the assurance that their identity would be protected. The respondents agreed to fill out a web questionnaire to provide demographic and psychometric information about themselves and to donate at least ten email messages.

Demographic traits cover basic demographic information about the author, namely age, gender, and level of education. The psychometric traits of the Arabic collection are based on a customized version of the short Eysenck Personality Questionnaire Revised (EPQR-S) (Francis et al., 2006), consisting of 51 questions. These questions aim to identify four psychometric traits: extraversion, lie (or social desirability), neuroticism (or emotionality), and psychoticism (or toughmindedness).

After completing the questionnaire, the writers either composed new email messages which they then sent to their recipients and also forwarded to the data collection email address, or directly forwarded previously sent emails, e.g. from their email client "SentBox".

We collected at least 10 emails from each author, for a total number of 2000 words per author in the Arabic data set. Research has shown that the more complex morphology of Arabic (combined with a rich vocabulary) leads to a higher degree of inherent sparseness in Arabic data compared to similar English data. This suggests that larger amounts of data are needed for statistical Natural Language Processing (NLP) applications in Arabic (Goweder and Roeck, 2001). Therefore, while the minimum was set at 1000 words per writer for the English data, it was 2000 words per writer for the Arabic data.

3.2 Data Validation

The email messages were 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. As with any data collection of email, plagiarism and copying were issues that required careful checking of all the data received and we developed a plagiarism detecter to reject emails which had already been submitted.

For both the English and the Arabic data collections, a minimum number of 5 lines per email had been set. Because Arabic writers often do not use new lines or new paragraphs, this had to be measured visually on the screen for the Arabic data.

In summary, the final Arabic data set contains a combined total of 8028 email messages, from 1030 writers who met the following criteria: 1) a valid questionnaire was received for each author; 2) there are at least 5 valid email messages for the author; and 3) the total word count for that author's valid email messages is at least 2000 words.

4 System Description

Figure 1 gives a high-level overview of the TAT system. The system includes several data repositories and a number of components for deriving features and for building classifiers. While the current focus is on processing email input in English and in Arabic, the underlying processing architecture is language independent and will be extended to other types of documents and to other languages.

The modular processing architecture is organized around a chain of processing modules. This chain allows flexible experimentation with various processing modules. At the same time, it provides a robust software framework that promotes reuse and supports flexible deployment options by connecting specific modules together for the task at hand.

Each processing module consumes objects from its input, such as documents, and emits objects containing the analysis of the input objects. The analysis of a document is represented in stand-off anno-

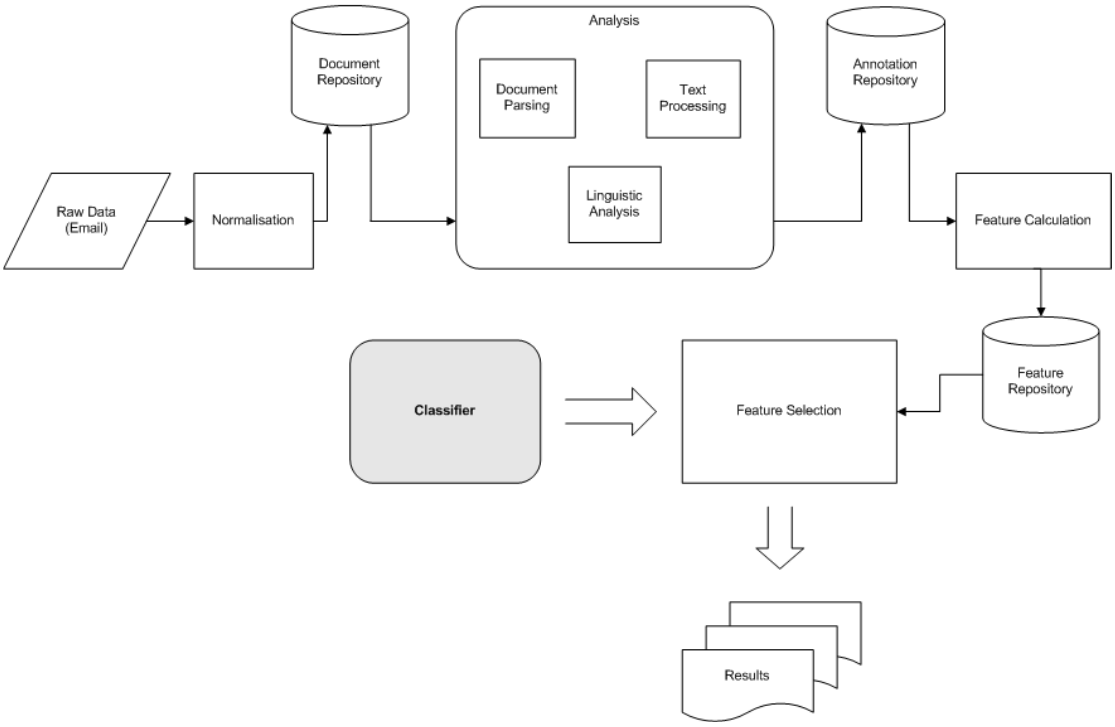

Figure 1: TAT System Diagram

tations and saved in a common structure called the Annotation Repository.

The process is data-driven in the sense that the output of each processing module depends on its input rather than on the way the module is combined with other modules in the chain. This enables processing modules to be reused in different processing chains and in different control environments as long as the input requirements are met.

5 Arabic Text processing

After the initial data collection and validation, several processing steps are needed, including character encoding normalisation, document structure parsing, text processing and linguistic analysis. The results of this processing provides the input to the feature extraction phase. A more extensive description of the processing steps for the English data can be found in (Estival et al., 2007).

Arabic emails present a number of challenges for NLP, including different ways of writing Arabic in Latin script (so-called “franco-arabic”); typical spelling variants in the Egyptian dialect of Arabic and possible spelling normalisation; morphology; English loanwords, spelling errors, and typos.

The Arabic processing modules perform the following functions:

1. Language and character-set identification. The language is tagged as either English or Arabic and the script as either “roman” or “arabic”, using character-based language models.
2. Document structure parsing. This stage distinguishes the text of the email written by the author from other types of elements (adverts, signatures, reply lines, or quoted text).
3. Tokenisation. The input text is split into paragraphs, sentences and words. Words are mainly strings of alphanumeric characters, with a few major exceptions for transliterated Arabic: some two character sequences, e.g. “ 3’ ”, can indicate a single Arabic glyph, namely غ.
4. Character set normalisation. In order to achieve a normalised, unambiguous input for Arabic and Latin scripted texts, words are converted to ASCII only characters using the Buckwalter transliteration scheme (Buckwalter, 2000; Buckwalter, 2002). The Buckwalter scheme

is very commonly used in NLP for Arabic because it represents Arabic orthography strictly on a one-to-one basis (unlike common romanization schemes that add morphological information not actually expressed in the Arabic script).

5. Spelling normalisation. Informal written Arabic often contains non-standard spellings (Buckwalter, 2004; Goweder and Roeck, 2001). We have therefore normalised the spelling based on common spelling variations in Egyptian Arabic. Examples include word final ى becoming ي or word final خ becoming ة. Also, depending on the context, ء can take different chairs or appear by itself on the line. We normalise the hamza chair to أ.
6. Morphological analysis. By comparing the normalised version of a given word with dictionaries of prefixes, suffixes and clitics, linguistic features such as number and person are added, and the remainder of the word is tagged as a stem. The root letters of the stem are then predicted using simple linguistic heuristics, e.g. long vowels are less likely to be root letters.
7. Lexicon taggers. The following word classes are currently tagged: conjunctions, prepositions, pronouns, discourse particles, interrogative pronouns, English loan words, colloquial Egyptian words, and frequent roots.
8. Named Entity Recognition. Named entities which are not language-dependent are tagged. These include numeric dates, numeric times, phone numbers, email addresses and URLs.

6 Experimental setup

Recent years have seen an exponential increase in the use of statistical language processing techniques for a wide range of tasks. In particular, text and document classification problems greatly benefit from statistical approaches. Author profiling can be viewed as a type of document classification task, where the classes correspond to traits of the authors. These traits are arranged along various dimensions and the different options for each dimension are mutually exclusive. For example “male” and “female” are the possibilities for the gender dimension. For each dimension, the email and questionnaire data are used to construct classifiers, using a range of Machine Learning (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 7 are on the entire data set. Once the best combination of ML classifiers, parameters and feature selection has been determined during development, that model is used to classify the test data to evaluate the performance for a given trait.

6.1 Traits and classes

We distinguish three different demographic and four psychometric traits in the experiments presented in this paper, namely age, gender, and level of education for the demographic traits, and extraversion, lie, neuroticism, and psychoticism for the psychometric traits (Eysenck and Eysenck, 1975). This information is extracted from the questionnaires.

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.

A major difference with regard to the psychometric traits described in (Estival et al., 2007) lies in the fact that we used two different questionnaires for English and for Arabic. The International Personality Item Pool questionnaire (IPIP) (Buchanan et al., 2005), which was used for collecting the English data, yields five psychometric traits: agreeableness, extraversion, neuroticism, conscientiousness, and openness (also referred to as the “Big Five”) (Norman, 1963). However the questionnaire for the Arabic data collection was based on the short form of the revised Eysenck Personality Questionnaire, the EPQR-S first developed for a study in Germany (Francis et al., 2006), and the adaptation of the EPQ for Arabic (Abdel-Khalek and Eysenenck, 1983). The EPQ (Eysenck and Eysenck, 1975) aims to analyze personality along four traits, namely extraversion, neuroticism, psychoticism, and lie. This entails that the Arabic results for the psychometric traits are not directly comparable to the English ones.

Demographics			
Age:	Gender:	Level of education:	
<25 (691)	Male (728)	No tertiary edu. (970)	
25 to 35 (236)	Female (302)	Some tert. edu. (60)	
>35 (103)			
Psychometrics			
Extraversion:	Lie:	Neuroticism:	Psychoticism:
low	low	low	low
medium	medium	medium	medium
high	high	high	high

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

6.2 Features

For each document, a feature vector is calculated. A feature is typically 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. Advert).

For the Arabic data, 518 features were calculated. These were divided into several subgroups shown in Table 3. The main purpose of the groupings was to make more informed choices during the feature selection stage and to facilitate experimentation with various combinations of feature groups.

Character-level features cover features such as the frequency of punctuation characters or word length. Arabic-specific character features include information on spelling normalisation and Arabic special characters. Morphological features mainly encode information on number, person and gender markers, such as clitics or suffixes. Lexical features include certain POS tags (e.g. preposition) and whether a word is an English loanword or specific to the Egyptian dialect.

6.3 Classification algorithms and feature selection

Classifiers for predicting author traits from the linguistic features were trained using the WEKA toolkit (Witten and Frank, 2005). During training, classifiers are created by the selection of sets of features for each 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 tested 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)), 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).

7 Results and discussion

The results shown here were computed on the Arabic email data set described in Section 3 using the different classifiers and general setup introduced in Section 6. Table 4 shows the results on all seven traits (demographic and psychometric), including the respective baseline associated with each separate classification task. We also state which settings (ML algorithm, feature set, and feature selection) were used to achieve the results reported. Education and gender are both binary classification tasks, whereas age uses three classes. All the psychometric traits are divided into three classes (see Section 6.1 and Table 2 for details on the exact split).

The results show that for six out of the seven traits

Feature group	Description
all	all features for Arabic
arabicNamedEntities	language independent named entities, such as URLs
arabicChar	Arabic character features
arabicMorphological	Arabic morphology features
arabicLexical	Arabic lexicon and word features

Table 3: Feature groups for Arabic.

Trait	ML algorithm	Features	Feature Sel.	Baseline	Results
Age:	Bagging	all except arabicLexical	InfoGain	70.09	72.10
Gender:	SMO	all	None	72.16	81.15
Education:	Bagging	all	InfoGain	93.62	93.66
Extraversion:	SMO	all except arabicMorphological	None	48.27	54.35
Lie:	Bagging	all	InfoGain	40.41	52.30
Neuroticism:	Bagging	all	InfoGain	43.42	54.93
Psychoticism:	Bagging	all	InfoGain	49.39	56.98

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

tested for, classification is significantly[1] improved over the baseline. For education, virtually no improvement can be seen which is due to the extremely skewed data (as indicated by the very high baseline of 93.62%). Even though the baselines for the other demographic traits are also quite high, our system still achieves a better classification accuracy for age and gender than the majority baseline.

The better result for gender can in part be explained by the fact that gender is morphologically marked in Arabic. One of the relevant constructions with respect to identifying an author's gender are predicative sentences with first person subjects. For example, in the Arabic equivalent of “I am happy”, *happy* is morphologically marked as either feminine or masculine. Since our features include morphological information, our classifier detects gender differences very accurately. A more detailed analysis of the effects of each feature group on the prediction of gender (shown in Table 5) reveals that lexical features are also of great assistance.

Table 6 shows the results for English and Arabic demographic traits that are directly comparable. This seems to confirm previous results showing that predicting author traits for Arabic is likely to be more difficult than for English. One should not forget, however, that the baselines for all Arabic demographic traits are extremely high which means that little data is available for the minority classes.

For the psychometric traits, we achieve similar improvements over the baseline as for English. This is particularly encouraging, as most research on Arabic author attribution has shown results for Arabic to be lower than for English. It seems that predicting a profile rather than an identity might be advantageous for Arabic, or at least a viable back-off option.

From our results, two ML algorithms emerge as best performing for all traits examined, namely SMO and Bagging. Bagging seems to profit from feature selection whereas the Support Vector Machine based SMO algorithm does not show additional improvements when combined with feature selection. This differs slightly from our results with the English data set, where no clear conclusion could be drawn with regard to the usefulness of feature selection for different algorithms.

An analysis of the results shows that the highest accuracy is achieved by including all available features, with the exclusion of a single feature group for age (arabicLexical) and for extraversion (arabicMorphological).

[1] Significance was tested using a χ^2 test with p=0.01.

Feature Group	Best Accuracy	Decrease in Accuracy
all	81.15	0.00
all except arabicNamedEntities	80.79	-0.36
all except arabicMorphological	80.19	-0.96
all except arabicChar	79.99	-1.16
all except arabicLexical	77.44	-3.71

Table 5: Contribution of feature groups to improvements in gender prediction.

Trait	English email data		Arabic email data	
	Accuracy	Improvement over Baseline	Accuracy	Improvement over Baseline
Age	56.46	+17.03	72.10	+02.01
Gender	69.26	+14.78	81.15	+08.99
Education	79.92	+21.14	93.66	+00.04

Table 6: Comparison of results for demographic traits for English and Arabic.

8 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 marketing intelligence to sentiment analysis. The results presented in this paper were conducted on the Arabic subset of the email data we have collected (containing approximately 8028 emails).

The experiments reported here were aimed at discovering how well a range of ML algorithms perform on our data set for three demographic and four psychometric author traits. Our results support the conclusions drawn in (Estival et al., 2007) 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.

Future work will include the extension of the TAT to other document types and other languages. For Arabic text processing, it might be fruitful to investigate a more sophisticated analysis of words into roots (Darwish, 2002; de Roeck and Al-Fares, 2000).

Acknowledgements

This research was carried out within the framework of a BAA grant (IS-QD-2467). We would like to thank Judith Bishop for her important contribution to the project in successfully managing the data collection.

References

Ahmed Abbasi and Hsinchun Chen. 2005a. 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.

Ahmed Abbasi and Hsinchun Chen. 2005b. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Ahmed Abdel-Khalek and Sybil Eysenenck. 1983. A cross-cultural study of personality: Egypt and England. *Research in Behaviour and Personality*, 3:215–226.

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.

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.

Tim Buckwalter. 2000. Arabic transliteration. http://www.qamus.org/transliteration.htm.

Tim Buckwalter. 2002. Arabic morphological analyzer. Linguistic Data Consortium (LDC2002L49).

Tim Buckwalter. 2004. Issues in Arabic orthography and morphology analysis. In *Proceedings of the COLING 2004 Workshop on computational approaches to Arabic script-based languages*, Geneva.

William Cohen. 1995. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123.

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.

Kareem Darwish. 2002. Building a shallow Arabic morphological analyzer in one day. In *Proceedings of ACL 2002 Workshop on computational approaches to Semitic languages*, Philadelphia.

Anne de Roeck and Waleed Al-Fares. 2000. A morphologically sensitive clustering algorithm for identifying Arabic roots. In *Proceedings of ACL 2000*, Toulouse.

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.

Olivier de Vel. 2000. Mining e-mail authorship. In *Proceedings of the Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000)*, Boston.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)*, pages ??–??, Melbourne.

Hans Eysenck and Sybil Eysenck. 1975. *Manual of the Eysenck Personality Questionnaire*. Hooder and Stoughton Educational, London.

Leslie Francis, Christopher Lewis, and Hans-Georg Ziebertz. 2006. The short-form revised Eysenenck personality questionnaire (EPQR-S): A German edition. *Social Behaviour and Personality*, 34(2):197–204.

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.

Abduelbaset Goweder and Anne De Roeck. 2001. Assessment of a significant Arabic corpus. In *Proceedings of the ACL/EACL Workshop on Arabic Language Processing: Status and Prospects*, Toulouse.

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, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. 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 209–217. Springer.

Gerrard Ledger and Thomas Merriam. 1994. Shakespeare, Fletcher, and The Two Noble Kinsmen. *Literary and Linguistic Computing*, 9(3):235–248.

F. Mosteller and D.L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Series in behavioral science: Quantitative methods edition. Addison-Wesley.

Warren T. Norman. 1963. Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, 66:574–583.

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.

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, 2nd 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.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *JASIST*, 57(3):378–393.