# Automatically Profiling the Author of an Anonymous Text

**Shlomo Argamon**, Illinois Institute of Technology
**Moshe Koppel**, Bar-Ilan University
**James W. Pennebaker**, University of Texas at Austin
**Jonathan Schler**, Bar-Ilan University

## 1. Introduction

Imagine that you have been given an important text of unknown authorship, and wish to know as much as possible about the unknown author (demographics, personality, cultural background, etc.), just by analyzing the given text. This *authorship profiling* problem is of growing importance in the current global information environment – applications abound in forensics, security, and commercial settings. For example, authorship profiling can help police identify characteristics of the perpetrator of a crime when there are too few (or too many) specific suspects to consider. Similarly, large corporations may be interested in knowing what types of people like or dislike their products, based on analysis of blogs and online product reviews. The question we therefore ask is: How much can we discern about the author of a text simply by analyzing the text itself? It turns out that, with varying degrees of accuracy, we can say a great deal indeed.

Unlike the problem of authorship *attribution* (determining the author of a text from a given candidate set), discussed recently in these pages by Li, Zheng, and Chen (2006), authorship profiling does not begin with a set of writing samples from known candidate authors. Instead, we exploit the sociolinguistic observation that different groups of people speaking or writing in a particular genre and in a particular language use that language differently (Chambers et al. 2004). That is, they vary in how often they use certain words or syntactic constructions (in addition to variation in, e.g., pronunciation or intonation). The particular profile dimensions we consider here are author *gender* (Argamon et al. 2003), *age* (Koppel et al. 2006), *native language* (Koppel et al. 2005) and *personality* (Pennebaker, Mehl, & Niederhoffer, 2003).

## 2. Text Categorization

Our approach to authorship profiling is to apply machine learning to text categorization (Sebastiani 2002). The process is as follows (see Figure 1): First, we take a given corpus of training documents, each labeled according to its category for a

particular profiling dimension. For example, when addressing classification by author gender, training documents are labeled as either ‘male’ or ‘female’. Each document is then processed to produce a numerical vector, each of whose elements represents some feature of the text that might help discriminate the relevant categories. A machine learning method then computes a classifier that, to the extent possible, classifies the training examples correctly. Finally, the predictive power of the classifier is tested on out-of-training data.

Essentially the same paradigm can be used for authorship attribution (Li, Zheng, and Chen 2006), where the training texts are known writings of given candidate authors. Text categorization methods within this paradigm have also been extensively applied to classifying documents by their *topic*. As we will describe, the key difference when classifying documents by authorial character is what features are used to represent the texts.

In the rest of the article, we first outline the kinds of text features that we find most useful for authorship profiling, and then describe the learning algorithms that we use for learning text classifiers. After that, we present experimental results for authorship profiling, analyzing which specific features prove to be the most effective discriminators for each problem.

## 3. **Features**

There are two basic types of features that can be used for authorship profiling: content-based features and style-based features (Campbell & Pennebaker, 2003). This reflects the fact that different populations might tend to write about different topics as well as to express themselves differently about the same topic. We shall consider these types separately, beginning with style-based features.

Many different types of features have been considered as possible markers of textual style including lexical, syntactic, and vocabulary complexity-based features. For special cases, other feature types may be considered, such as grammatical or orthographic errors in unedited text, or morphological features for languages with especially rich morphological structure. However, the preponderance of evidence

suggests that the most consistently effective features over a wide variety of authorship problems are function words and individual parts-of-speech.

In this work, we used a novel feature set that naturally subsumes both function words and the parts-of-speech known to be useful in stylistics. Systemic Functional Linguistics (Halliday & Matthiessen 2003) provides taxonomies describing meaningful distinctions among various function words and parts-of-speech (see Figure 2). We represent such taxonomies as trees whose roots are labeled by sets of parts-of-speech (articles, auxiliary verbs, conjunctions, prepositions, pronouns). Each node's children are labeled by meaningful subclasses of the parent node (such as the various sorts of personal pronouns). This bottoms out at the leaves, which are labeled by sets of individual words. The set of function words corresponds to the set of all the articles, auxiliary verbs, conjunctions, prepositions and pronouns that appear as leaves in these trees. Our feature set is composed of, for each node in these taxonomic trees, the frequency of the node's occurrence in a text normalized by the number of words in the text.

In addition to the style-related features just described, we also consider content-based features, namely, individual words. In order to keep the number of features reasonably small, we consider just the 1000 words that appear sufficiently frequently in the corpus and that discriminate best between the classes of interest (as determined by applying the "information-gain" measure (Mitchell, 1999) to a holdout set broken off of our training data).

We should note that the use of content-based features for authorship studies can be problematic. While it is plausible that style-based markers can truly distinguish one class of authors from another, one must be wary that content markers might just be artifacts of a particular writing situation or experimental setup and might thus produce overly optimistic results that will not be borne out in real-life applications. We are therefore careful to distinguish results that exploit content-based features from those that do not.

## 4. **Learning Methods**

Whatever features are used in a particular experiment, we represent a document as a numerical vector $X = (x_1,..,x_i,\ldots,x_n)$, where $n$ is the number of features and $x_i$ is the relative frequency of feature $i$ in the document. Once labeled training documents have been represented in this way, we can apply machine-learning algorithms to learn classifiers that assign new documents to categories. The most effective multi-class (i.e., more than two classes) classifiers for authorship studies all share the same structure: We learn a weight vector $W^j = (w_1^j,\ldots,w_i^j,\ldots,w_n^j)$ for each category $c_j$ and then assign a document, $X$, to the class for which the inner product $W^j * X$ is maximal. There are a number of effective algorithms for learning the weight vectors (e.g., Crammer & Singer 2003; Genkin et al. 2006, Schler et al. 2006). In this paper, we use as our learning algorithm Bayesian Multinomial Regression (BMR) (Genkin et al. 2006) which we have found to be both efficient and accurate. BMR is a probabilistically well-founded multivariate variant of logistic regression which is resistant to overfitting; it has been shown to be effective for text classification and related problems.

## 5. **Experimental Setup**

We will consider four profiling problems: determining the author's gender, age, native language, and neuroticism level. Ideally, we would prefer to use a single corpus for all these problems but, unfortunately, there is no single corpus in which the documents are labeled for all four issues we consider. Thus we use three separate corpora (age and gender experiments use the same corpus), as will be described below. For each of the three feature sets – stylistic features only, content features only, and both – we run ten-fold cross-validation tests to test the extent to which each profiling problem is solvable. We will also present the most discriminating features for each category within each of the four problems.

We now consider in turn each of our four profiling problems.

### *5.1 Gender*

Our corpus (first described in (Schler et al. 2006)) for both gender and age consists of the full set of postings of 19,320 blog authors (each text is the full set of posts by a

given author) writing in English . The (self-reported) age and gender of each author is known and for each age interval the corpus includes an equal number of male and female authors. The texts range in length from several hundred to tens of thousands of words, with a mean length of 7250 words per author.

Classification results for gender are shown in the first line of Table 1. As is evident, all feature sets give effective classification, while the content features are slightly better than style features.

In Table 2, we show the most discriminating style and content features, respectively, for each experiment. The features are ranked using the information-gain measure for continuous features as described in (Mitchell 1999). For the parts-of-speech, we avoid repetition by listing only those parts-of-speech for which no subset (other than individual words) or superset has already been listed.

As can clearly be seen in Table 2, the style features that prove to be most useful for gender discrimination are determiners and prepositions (markers of male writing) and pronouns (markers of female writing). (Note that for the case of gender, as with all the other problems we consider here, the most discriminating style features include parts-of-speech and not only function words. Restricting our feature set to function words alone diminishes accuracy in each of our experiments by 5-10%.) The content features that prove to be most useful for gender discrimination are words related to technology (male) and words related to personal life or relationships (female).

Keep in mind that these content features are appropriate for gender classification on this specific collection of blogs; for other types of text, other content features may be more appropriate. This genre-dependency holds also for other authorship profiling dimensions. Nevertheless, earlier studies (Argamon et al. 2003) on author gender in both fiction and non-fiction have shown that similar style features to those shown here are appropriate for other types of text as well.

*5.2 Age*

Based on each blogger's reported age, we label each blog in our corpus as belonging to one of three age groups: 13-17 (42.7%), 23-27 (41.9%) and 33-47 (15.5%).

Intermediate age groups were removed to avoid ambiguity since many of the blogs were written over a period of several years. (This might make our version of the problem slightly easier than that which might be encountered in real-life.) Our objective is to identify to which of these three age intervals an anonymous author belongs.

Results for age are shown in the second line of Table 1. Here too, we see that both style and content features give us over 76.1% accuracy for this three-way classification problem, while the baseline majority-class classifier would give 42.7%.

The style features that prove to be most useful for discrimination are contractions without apostrophes (younger writing), and determiners and prepositions (older writing). Note that the strongest style features for 20s and 30s are identical; they are those that distinguish both of them from teenagers. The content features that prove to be most useful for discrimination are words related to school and mood for teens, to work and social life for twenties, and to family life for thirties.

*5.3 Native Language*

We use the International Corpus of Learner English (ICLE) [http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm], which was assembled for the precise purpose of studying the English writing of non-native English speakers from a variety of countries. All the writers in the corpus are university students (mostly in their third or fourth year) studying English as a second language. All are roughly the same age (in their twenties) and are assigned to the same proficiency level in English. We consider five sub-corpora, from Russia, the Czech Republic, Bulgaria, France, and Spain. To balance the corpus, we took 258 authors from each sub-corpus (randomly discarding any surplus). All texts in the resulting corpus are between 579 and 846 words long. Our objective is to determine which of the five languages is the native tongue of an anonymous author writing in English.

Results are shown in the third line of Table 1. Both style and content features give results far above the baseline of 20% (assigning any constant class).

Some consistent patterns of usage can be seen in the style features. For example, as might be expected, native speakers of Slavic languages (Russian, Bulgarian, Czech) tend to omit the definite article *the* which does not exist in these languages. (Since we list only features that are over-represented in a given class, this feature is seen by examining the list of features for Spanish. Indeed, many of the most discriminating features are those that are *under*-represented for particular languages.) Furthermore, those words with commonly used analogs in a given language are used with greater frequency by native speakers of that language, such as *indeed* (French), *over* (Russian), and *however* (Bulgarian).

Elsewhere (Koppel et al. 2005), we have shown that specifically for determining native language, features that measure stylistic idiosyncrasies and errors are particularly useful. For example, Romance language speakers often use the vowel 'o' where standard English specifies another vowel (e.g., *outhor* for *author*). Using such features together with the style features considered in this paper yields classification accuracy of over 80% for this task.

Regarding content words, we see that, in this corpus, speakers of different native languages use certain words more than others. It should be noted, however, that unlike the text collections used in the other experiments described in this paper, writers in the learner corpus did not freely choose their writing topics. Rather, they were assigned topics from an ICLE standard set of possible writing topics. Since the precise instructions given to students in different countries may have varied somewhat, differences in content word usage here are most likely artifacts of the experimental setup.

*5.4 Personality*

We use essays written by psychology undergraduates at the University of Texas at Austin as part of their course requirements. Students were instructed to write a short "stream of consciousness" essay wherein they tracked their thoughts and feelings over a 20-minute free-writing period. The essays range in length from 251 to 1951 words. Each writer also filled out a questionnaire testing for the "Big Five" personality dimensions: *neuroticism*, *extraversion*, *openness*, *conscientiousness*, and

*agreeableness* (John et al 1991). To illustrate personality profiling, we consider just the dimension of neuroticism (roughly: tendency to worry); methods and results for other personality factors are qualitatively similar. To formulate this as a classification problem, we define 'positive' examples to be the participants with neuroticism scores in the upper third, and 'negative' examples to be those with scores in the lowest third. The rest of the data are ignored; the final corpus consists of 198 examples.

Results are shown in the fourth line of Table 1. Significantly, style features give a great deal of information about personality. An accuracy rate of 65.7% in detecting neuroticism (where 50% is chance) is surprisingly high. Unlike sex or age, there are few visible markers of being high in anxiety. Independent studies of close friends who attempt to guess people's neuroticism classification averages 69% - even among people who have known each other for several years (Vazire, 2006).

As Table 2 shows, the most discriminating style features indicate that neurotics tend to refer to themselves, use pronouns for subjects rather than as objects in a clause and reflexive pronouns, and consider explicitly who benefits from some action (through prepositional phrases involving, e.g., "for" and "in order to"); non-neurotics, on the other hand, tend to be less concrete and to use less precise specification of objects or events (determiners and adjectives such as "a" or "little") and show more concern with how things are or should be done (via prepositions such as "by" or "with" and modals such as "ought to" or "should").

In fact, classifiers learned using only the ten style features shown in Table 2 give classification accuracy of 63.6%. More surprisingly, although the results in Table 1 indicate that content words overall are useless for classifying texts by neuroticism, using as features the ten most informative content features (those in Table 2) gives an accuracy of 68.2%. Apparently, the vast majority of content is irrelevant to this classification problem and masks a small number of features involving worry about personal problems (neurotics) and relaxation activities (non-neurotics) that are quite useful for this task.

## 6. Conclusions

Accurate profiling of the demographics, background, and personality of an unknown author is a task of growing importance for national security, criminal investigations, and market research. We have shown how the right combination of linguistic features and machine learning methods enables an automated system to effectively determine several such aspects of an anonymous author; it is likely that other important profile components (such as educational background or other personality components) can also be extracted using such techniques, given appropriate training material. An important open research question, however, is the extent to which variation in genre and language might affect the nature of the models that can be used to solve various aspects of the profiling problem.

## 7. Bibliography


1. Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, **23**(*3*).
2. Chambers, J. K., Trudgill, P., and Schilling-Estes, N., eds. (2004). *The Handbook Of Language Variation And Change* (London: Blackwell).
3. Crammer, K. and Singer, Y. (2003). Ultraconservative Online Algorithms for Multiclass Problems. *Journal of Machine Learning Research*, **3**:951—991.
4. Genkin, A., Lewis, D. D., and Madigan, D. (2006). Large-scale Bayesian logistic regression for text categorization. *Technometrics* (to appear).
5. Halliday, M. A. K., and Matthiessen, C. (2004). *An Introduction To Functional Grammar* (3rd ed.) (London: Arnold).
6. Koppel, M., Schler, J., and Zigdon, K. (2005), Determining an Author's Native Language by Mining a Text for Errors (short paper), *Proceedings of KDD*, Chicago IL, August 2005.
7. Koppel, M., Schler, J., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. Presented at *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford, CA, March 2006.
8. Li, J., Zheng, R., and Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM* **49**:4 (Apr. 2006), pp. 76-82.
9. Mitchell, T. (1999). *Machine Learning*. (New York: McGraw-Hill)

10. Pennebaker, J.W., Mehl, M.R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577.
11. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1).
12. Vazire, S. (2006). *The person from the inside and the outside*. Unpublished doctoral dissertation. The University of Texas, Austin, TX, USA.

**Table 1.** Classification accuracy for profiling problems using different feature sets.

| | **Baseline** | **Style** | **Content** | **Style+Content** |
|---|---|---|---|---|
| **Gender (2 classes)** | *50.0* | 72.0 | 75.1 | **76.1** |
| **Age (3 classes)** | *42.7* | 66.9 | 75.5 | **77.7** |
| **Language (5 classes)** | *20.0* | 65.1 | **82.3** | 79.3 |
| **Neuroticism (2 classes)** | *50.0* | **65.7** | 53.0 | 63.1 |

**Table 2.** Most important Style and Content features (by information gain) for each class of texts in each profiling problem.

| **Class** | **Style Features** | Content Features |
|---|---|---|
| Female | **personal pronoun, *I*, *me*, *him*, *my*** | ***cute, love, boyfriend, mom, feel*** |
| Male | **determiner, *the*, *of*, preposition-matter, *as*** | *system, software, game, based, site* |
| Teens | ***im*, *so*, *thats*, *dont*, *cant*** | *haha, school, lol, wanna, bored* |
| Twenties | **preposition, determiner, *of*, *the*, *in*** | *apartment, office, work, job, bar* |
| Thirties+ | **preposition, *the*, determiner, *of*, *in*** | *years, wife, husband, daughter, children* |
| Bulgarian | **conjunction-extension, pronoun-interactant,** *however*, **pronoun-conscious,** *and* | *bulgaria, university, imagination, bulgarian, theoretical* |
| Czech | **personal pronoun,** *usually*, *did*, *not*, *very* | *czech, republic, able, care, started* |
| French | *indeed*, **conjunction-elaboration,** *will*, **auxverb-future, auxverb-probability** | *identity, europe, european, nation, gap* |
| Russian | *can't*, *i*, *can*, *over*, *every* | *russia, russian, crimes, moscow, crime* |
| Spanish | **determiner-specific**, *this, going_to, because, although* | *spain, restoration, comedy, related, hardcastle* |
| Neurotic | *myself*, **subject pronoun, reflexive pronoun, preposition-behalf, pronoun-speaker** | *put, feel, worry, says, hurt* |
| Non-neurotic | ***little*, auxverbs-obligation, nonspecific determiner, *up*, preposition-agent** | *reading, next, cool, tired, bed* |

**Figure 1.** Architecture for authorship profiling using machine learning. Documents labeled for a given authorship characteristic, such as gender (generically, "A" and "B"), are used as training data; they are linguistically processed and tagged and feature frequencies calculated, giving a numeric vector for each individual text, labeled with the text's correct authorship label. A machine learning method creates a classification model from this training that is then applied to vectors computed from unlabeled test documents – classification accuracy gives a measure of how effective the technique is, while the most significant features for classification give a rough characterization of the linguistic difference between given author types.

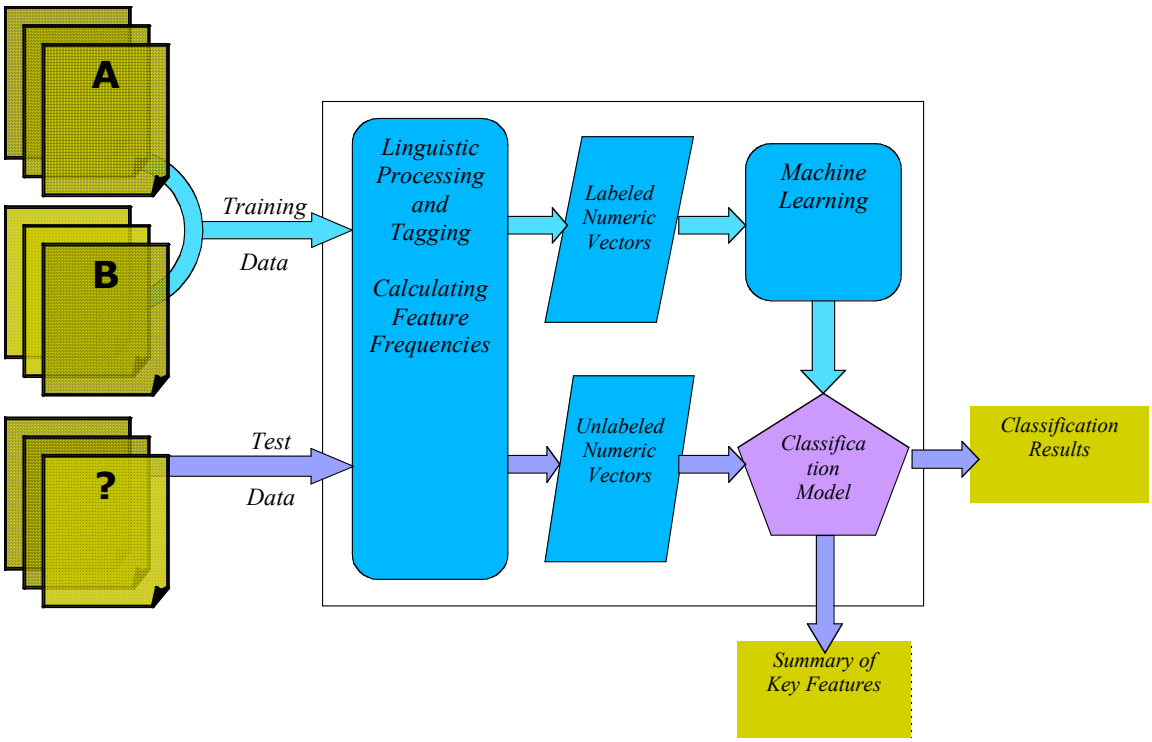

**Figure 2.** Two of the functional word taxonomies used in our system.