

Automatically Categorizing Written Texts by Author Gender

Moshe Koppel
Shlomo Argamon
Anat Rachel Shimoni

Dept. of Computer Science, Bar-Ilan University, Israel

Published in *Literary and Linguistic Computing* **17**(4), 2002

Presented by Afra Alishahi. February 2004

Gender-specific linguistic style

- Goal: to explore the possibility of automatically classifying formal written texts according to author gender.
- Previous work:
 - ▶ On spoken language, using intonational, phonological and conversational cues
 - ▶ On writing on more informal contexts, e.g., student essays or electronic communications

A text categorization problem?

Components of a text categorization system:

- Document representation
 - ▶ Selecting suitable features
 - ▶ Representing each text as a vector of frequencies
- Dimension reduction
 - ▶ Eliminating those features that are not relevant
- Learning method
 - ▶ Constructing a model for each category
- Testing protocol

A stylometry problem?

- Text categorization by topic
 - ▶ based on keywords which reflect a document's content
 - ▶ relatively large feature sets
 - ▶ machine learning algorithms

- Stylometry
 - ▶ based on content-independent features
 - ▶ hand-selected sets of lexical, syntactic or complexity-based features
 - ▶ statistical methods

A hybrid problem

- The current problem is different from text categorization – it does not depend on the content.
- It also differs from stylometry – there is no individual author whose style habits are exhibited in the text.
- The authors use ideas from both camps to accomplish the task.

The corpus

- The BNC includes 920 gender- and genre-labeled documents.
- In each sub-genre, $\min(\text{male}, \text{female})$ documents and a randomly selected equal number of documents from the other class are selected.
- The resulting corpus contains 566 documents.
- No single author has more than three documents in this corpus.
- The documents contain between 554 and 61,199 words with an average of about 34,320 words each.

Document representation

■ Features (total = 1081)

- ▶ 405 function words
- ▶ all the 76 single part of speech tags
- ▶ 100 most common ordered POS pairs
- ▶ 500 most common ordered POS triples

■ Vectors

- ▶ Each document is represented as a vector of length 1081
- ▶ Each entry represents the number of appearances of the feature in the document divided by document length
- ▶ Function word and POS pair frequencies were multiplied by 2
- ▶ POS triple frequencies were multiplied by 4

The learning method

- Goal: to find a linear separator between male-authored and female-authored documents:

- ▶ Searching for a weight vector w such that for each document x ,

$$w * x > T \iff x \text{ is authored by a female}$$

where T is a threshold value.

- A variant of the Exponential Gradient (EG) algorithm of (Kivinen & Warmuth 1997) is used.
- The learning algorithm itself is used for feature reduction: low-weighted features are eliminated.

The learning method (cont.)

■ The learning algorithm:

- ▶ $w^+ = \{1, 1, \dots, 1\}$, $w^- = \{-1, -1, \dots, -1\}$, $w = w^+ + w^-$
- ▶ $c(x) = 1$ if x is female-authored and $c(x) = 0$ otherwise.
- ▶ $s(w, x) = 1$ if $w * x > 0$ and $s(w, x) = 0$ otherwise.
- ▶ Iteratively update the weights for each training document:

$$w_i^+ \leftarrow w_i^+ (1 + \beta x_i)^{(c(x) - s(w, x))}$$

$$w_i^- \leftarrow w_i^- (1 + \beta x_i)^{(s(w, x) - c(x))}$$

- ▶ After all training documents have been used, they are randomly reordered and another cycle of updates is run.
- ▶ This continues until all training examples are correctly classified or until 100 consecutive cycles fail to improve the performance.

Results

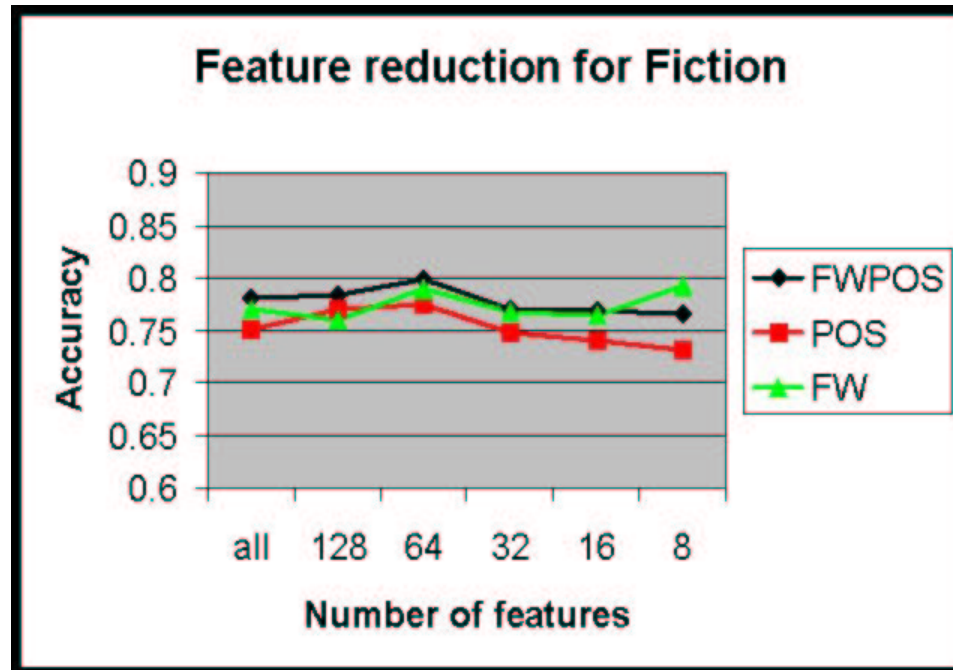
Accuracies for training/testing on all documents, fiction, and nonfiction:

Domain	FW	POS	FWPOS
All	73.7	70.5	77.3
Fiction	78.8	77.1	79.5
Non-fiction	68.5	67.2	82.6

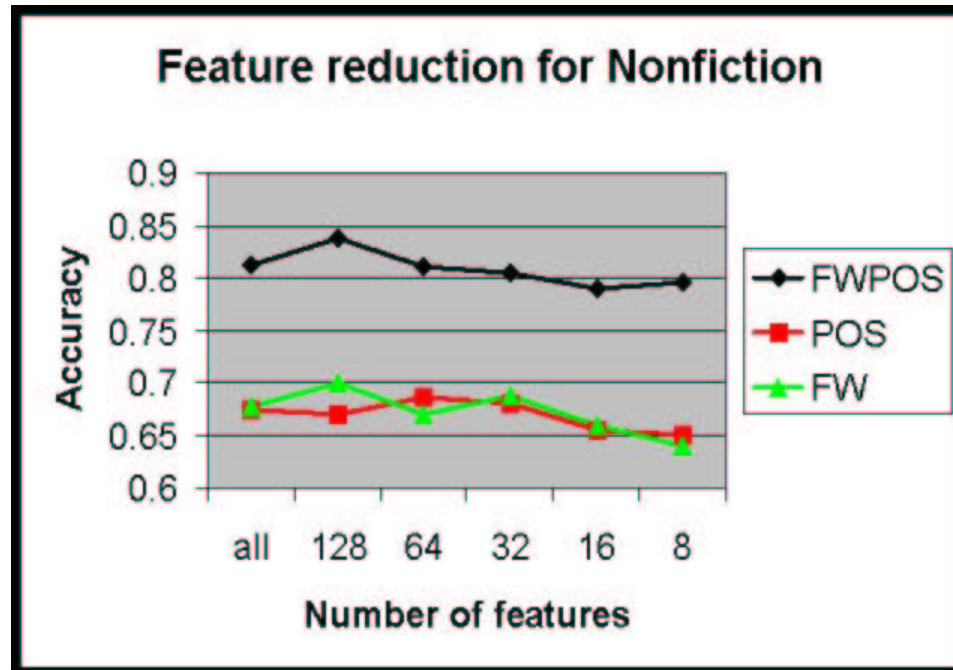
Feature reduction

- For each model obtained in a cross-validation trial, top 128 features were selected.
 - ▶ feature rank in a model is defined as the absolute value of its weight multiplied by its average frequency in the training set.
- The cross-validation trial was executed again, using the reduced set of features.
- The number of features was cut in half, and the above process was repeated, until only 8 features left on each side.

Feature reduction for fiction



Feature reduction for nonfiction



Feature analysis

- Function words for fiction:
 - ▶ Male features – *a, the, as*
 - ▶ Female features – *she, for, with, not*
- Function words for nonfiction:
 - ▶ Male features – *that, one*
 - ▶ Female features – *for, with, not, and, in*
- Parts of speech:
 - ▶ Male features – determiners, numbers, modifiers
 - ▶ Female features – negation, pronouns, certain prepositions

Categorization by genre

- Results of using the same system for distinguishing between the two genres fiction and nonfiction:

Feature set	Accuracy
FWPOS	98.2
POS	97.5
FW	97.9

- All misclassified nonfiction documents are biographical or diary-like works.

Conclusion

- Automatic text categorization techniques were used to infer the gender of the author of an unseen formal written document with approximately 80% accuracy.
- Best performance is achieved when both function words and parts-of-speech n-grams are used.
- A relatively small number of such features is required for reasonable categorization.
- The method works for other style-based categorization problems, e.g., distinguishing fiction from nonfiction.