

available at www.sciencedirect.comDigital
Investigationjournal homepage: www.elsevier.com/locate/diin

Author gender identification from text[☆]

Na Cheng, R. Chandramouli*, K.P. Subbalakshmi

Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA

ARTICLE INFO

Article history:

Received 1 July 2010

Received in revised form

30 March 2011

Accepted 24 April 2011

Keywords:

Gender identification

Text mining

Psycho-linguistic analysis

Logistic regression

Decision tree

Support vector machine

ABSTRACT

Text is still the most prevalent Internet media type. Examples of this include popular social networking applications such as Twitter, Craigslist, Facebook, etc. Other web applications such as e-mail, blog, chat rooms, etc. are also mostly text based. A question we address in this paper that deals with text based Internet forensics is the following: *given a short text document, can we identify if the author is a man or a woman?* This question is motivated by recent events where people faked their gender on the Internet. Note that this is different from the authorship attribution problem.

In this paper we investigate author gender identification for short length, multi-genre, content-free text, such as the ones found in many Internet applications. Fundamental questions we ask are: do men and women inherently use different classes of language styles? If this is true, what are good linguistic features that indicate gender? Based on research in human psychology, we propose 545 psycho-linguistic and gender-preferential cues along with stylometric features to build the feature space for this identification problem. Note that identifying the correct set of features that indicate gender is an open research problem. Three machine learning algorithms (support vector machine, Bayesian logistic regression and AdaBoost decision tree) are then designed for gender identification based on the proposed features. Extensive experiments on large text corpora (Reuters Corpus Volume 1 newsgroup data and Enron e-mail data) indicate an accuracy up to 85.1% in identifying the gender. Experiments also indicate that function words, word-based features and structural features are significant gender discriminators.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The rapid growth of the Internet has created myriad ways to share information across time and space. Online social networking (such as Twitter, Myspace, Facebook), e-commerce (such as eBay, Craigslist), usenet newsgroups, etc. are gaining more prominence. As of October 2009, the number of the Internet users is estimated to be 1.69 billion according to statistics evaluated by AMD (50x50).

However, this growth also encourages various kinds of misuses. Online communities are vulnerable to deceptive attacks, receiving false information, etc. The 2008 Annual Report of Internet Crime Complaint Center (IC3) states that there was a 33.1% increase in online crime in 2008. In order to mitigate this situation, homeland security and law enforcement agencies have launched projects to prevent deceptive attacks and track the identities of senders to protect against terrorism, child predators, etc.

[☆] Preliminary version of this work was presented in the IEEE Symposium on Computational Intelligence and Data Mining Conference, April 2009.

* Corresponding author. Tel.: +1 201 216 8642; fax: +1 201 216 8246.

E-mail addresses: ncheng@stevens.edu (N. Cheng), mouli@stevens.edu (R. Chandramouli), ksubbala@stevens.edu (K.P. Subbalakshmi).
1742-2876/\$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.diin.2011.04.002

Anonymity is a significant characteristic in online communities (Zheng et al., 2006; Abbasi & Chen, May 2006; Chen, 2005). People may not need to provide their true identity such as name, age, gender and address in cyberspace. In many misuses or crime cases, the perpetrators attempt to hide their addresses by using anonymous servers, and conceal their real identities to avoid being detected. Therefore, it becomes imperative to design an efficient method for identity tracing in cyberspace forensics. We are particularly interested in gender identification in our paper.

To illustrate the need to study the gender identification problem let us consider the recent “Myspace mom” case (www.foxnews.com). Lori Drew (female) and a few others pretended to be a teenage boy (called Josh) on Myspace and befriended Megan Meier. Megan and Josh began exchanging messages on Myspace for more than a month when Josh abruptly ended their friendship, telling Megan that she was cruel. This soon led to the suicide of Megan. This case clearly has implications for text based gender identification techniques on safeguarding children on the Internet. Missouri was the first state to enact anti-cyberbullying legislation after this case.

Psychology research suggests that one’s state of mind, such as physical/mental health and emotion, can be gauged by the words he/she uses (e.g., Pennebaker, 1995; Newman et al., 2003; Peng et al., 2003) suggest that each author has a unique stylistic tendency, and refer to this feature as the author profile. Using these textual traces, researchers have begun to use online stylometric analysis techniques (authorship identification) as a forensic identification tool. But note that an author could conceal his/her true name by changing it to a name from the opposite sex.

Related work on authorship identification (not gender identification) studies dates back to the 18th century when English logician Augustus de Morgan suggested that authorship might be settled by determining if one piece of text contained significantly longer words than another (Zheng et al., 2006). Well known authorship identification studies include the attribution of disputed Shakespearean works (e.g. Efron and Thisted, 1976; Lowe and Matthews, 1995; Merriam, 1996) and the attribution of the Federalist papers (Mosteller and Wallace, 1964; Holmes and Forsyth, 1995; Tweedie et al., 1996).

With the development of computers, stylometry has been widely accepted and has become dominant in identifying authorship. Over 1000 stylometric features have been proposed so far, including word- or character-based stylometric features (e.g. Yule, 1944; Holmes, 1992), function words (e.g. Mosteller and Wallace, 1984; Burrows, 1987), punctuation (e.g. Baayen et al., 2002), etc. Although authorship identification methods have achieved some degree of success in many literary and forensic applications as mentioned above, very limited studies have been undertaken specifically for online messages. There are also several other approaches to authorship identification. Statistical methods were widely used in earlier studies based on histograms of word-length distribution of various authors (Mendenhall, 1887), the Bayesian classifier (Mosteller and Wallace, 1984), principle component analysis (Burrows, 1987), cluster analysis (Holmes, 1992), etc. The advent of powerful computers instigated the

extensive use of machine learning techniques in authorship analysis, such as decision tree (Apte et al., 1998), neural networks (Tweedie et al., 1996), support vector machine (SVM) (Diederich et al., 2000), etc.

The problem we address in this paper—author gender identification—from short Internet text is different from other types of authorship identification/attribution problems, due to the following:

- gender identification is a higher level of abstraction; unlike authorship attribution the candidate set of authors is unavailable a priori
- the length of Internet text messages is usually small compared to traditional text documents such as books for which authorship attribution is mostly studied
- unlike traditional text documents special linguistic elements such as emoticons often appear in Internet texts
- the format or the structure of Internet texts may vary among different users and situations due to real-time constraints such as Internet chat, instant messaging, etc.

Fundamental questions we ask are: do men and women inherently use different classes of language styles? If this is true, what are reliable linguistic features that indicate gender?

Generally, men and women converse differently even though they technically speak the same language. Many researches have been studying the relationship between gender and language use (e.g., Online, 2002). Robin Lakoff’s *Language and Woman’s Place* (Lakoff, 1975) has influenced the research on language and gender since it was published in 1975. Lakoff presented a group of lexical, syntactic and pragmatic features, which would distinguish the language style of women, viz., the use of specialized vocabulary, expletives, tag questions, hedges and hypercorrect grammar. Mary Talbot indicates in her book (Talbot, 1998) that different patterns of language use reflect the social divisions on gender grounds. According to Talbot, women who are subordinate in status to men in work settings exhibit greater use of polite language. Mulac et al. (1990) and Mulac and Lundell (1994) introduced gender-linked effects by analyzing students’ impromptu essays, descriptions of photographs and problem-solving dyadic interactions between strangers. They also summarize more than 30 studies to show that gender differences exist in written communication in a wide variety of contexts as well as in face-to-face interaction (Mulac, 1998).

Note that researchers distinguish between sex and gender, since sex is biologically founded, whereas gender is socially constructed (Crawford, 1995). Hence, by birth one’s sex may be female or male, but they are taught by society how to fulfill their corresponding gender. It needs to be made clear that all men are not masculine and all women are not feminine. Our research is based on the concept of gender-related language, rather than sex-related. For consistency purpose, we still use male/female to denote masculine/feminine.

Relatively little work (Vel et al., 2002) has been done in gender identification from text. Therefore, in this work we propose robust classifiers based on content-free feature set to identify the gender of the author of short text messages typically found on the Internet. The contributions of this paper are:

- propose of several types of features as gender indicators
- design of a set of measures to infer the author’s gender from short messages
- design of classifiers and their parameter optimization

The rest of this paper is organized as follows: Section 2 presents the background, data pre-processing, feature set selection and the proposed gender identification methods. Section 3 presents extensive experiment results to evaluate the proposed features methods. Concluding remarks are presented in Section 4.

2. Author gender identification

2.1. Problem formulation

The author gender identification problem can be treated as a binary classification problem, i.e., given two classes {male, female}, assign an anonymous text message e to one of these classes:

$$e \in \begin{cases} \text{Class}_1 & \text{if the author of } e \text{ is male} \\ \text{Class}_2 & \text{if the author of } e \text{ is female} \end{cases} \quad (1)$$

To design a hypothesis test (1) we have to design a set of features that remain relatively constant for a large number of messages written by authors of the same gender. Once the feature space has been designed a given message e can be represented by a d -dimensional vector where d is the total number of features. Given a set of known pre-classified messages, a model (or classifier) is built which can then be used to determine the category of a given message.

Mathematically, we are interested in learning classifiers $y = f(x)$, from a set of training examples $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$. Let $X = \{x_i, i = 1, 2, \dots, N\}$ denote an instance set, where x_i is a d -dimensional vector $x_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$. Let $Y = \{y_i, i = 1, 2, \dots, N\}$ denote the label set, where $y_i \in \{+1, -1\}$ is class label encoding class₁ (-1) or class₂ (+1) and N is the number of examples in the dataset. In general, the gender identification process can be divided into four steps (see, Fig. 1)

- 1) collecting a suitable corpus of text messages to be the dataset;
- 2) identifying features that are significant indicators of gender;
- 3) extracting feature values from each message automatically;
- 4) building a classification model to identify the author’s gender of a candidate text message.

2.2. Dataset pre-processing

Among various types of Internet text messages, newsgroup messages make use of neutral descriptive language while more private personal e-mails better reflect the true character of the author. Therefore, we use these two extreme types of datasets in the classifier design.

2.2.1. Reuters newsgroup dataset

Reuters is the world’s largest international multimedia news agency, providing myriad news and mutual fund information available on Reuters.com, video, mobile, and interactive television platforms. Reuters Corpus Volume 1 (RCV1) is drawn

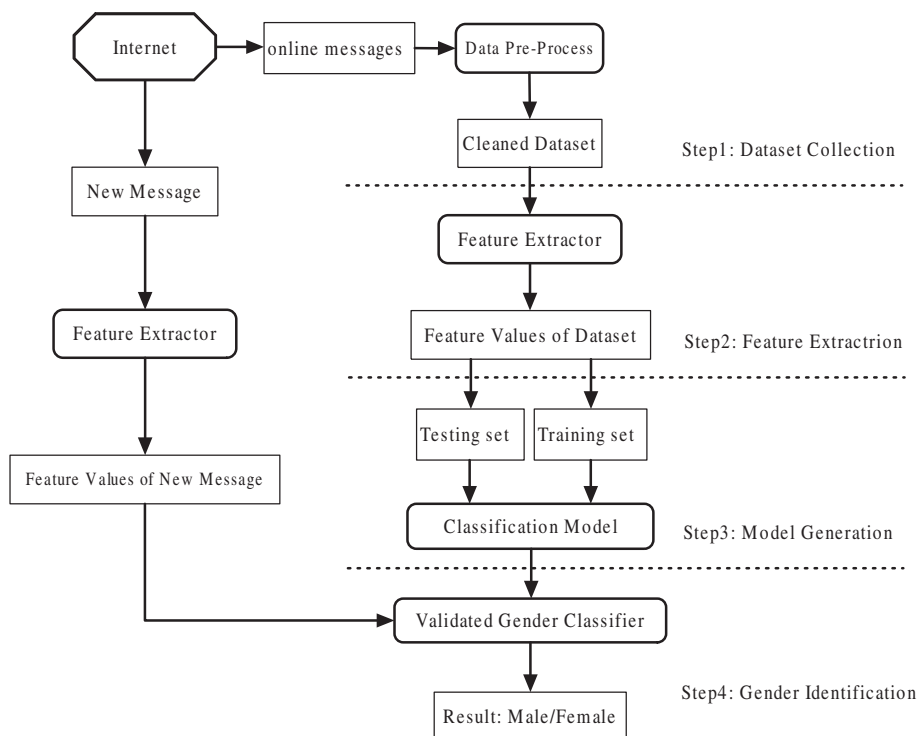


Fig. 1 – Gender identification process.

from one of those online databases (Online, 2000). This dataset consists of all English language stories produced by Reuters journalists between August 20, 1996 and August 19, 1997. The dataset is made available on two CD-ROMs and has been formatted in XML by Reuters, Ltd. in 2000, for research purposes. Both the archiving process and later preparation of the XML dataset involved substantial verification and validation of the content, attempts to remove spurious or duplicated documents, normalization of dateline and byline formats, addition of copyright statements, and so on. The stories cover a range of content typical of a large English language international newswire. They vary from a few hundred to several thousand words in length.

We retrieved the information of the authors, categorized the documents by the gender of the journalists, and discarded the documents written by authors whose gender could not be ascertained (e.g., neutral names). The text messages were then extracted by removing unnecessary information (such as date and time) and all XML formatting. Then we kept only the messages that contained more than 200 but less than 1000 words. Since most messages in the Reuters Corpus were newswire stories, these messages contained several quotes from others, which could bias the accuracy of the gender classifier. Therefore, we removed those messages that had too many quotations (") by setting the threshold of number of quotes/character counts to be 0.002. Table 1 summarizes the corpus used in our analysis.

2.2.2. Enron e-mail dataset

A suitable corpus of e-mails was generally lacking for gender identification research. This was due to privacy and other considerations. Therefore, we used the Enron e-mail dataset (Online, 2005). Enron, the energy company based in Houston, Texas went bankrupt in 2001 because of accounting fraud. During the process of investigation, the e-mails of employees were made public by the Federal Energy Regulatory Commission.

Several integrity problems with the Enron e-mail corpus have been corrected by a number of other researchers. Some email messages were deleted from the dataset as part of a redaction effort due to requests from affected employees. Invalid e-mail addresses were converted to something of the form `user@enron.com` whenever possible (i.e., recipient is specified in some parse-able format like "Doe, John" or "Mary K. Smith") and to `no_address@enron.com` when no recipient was specified.

The final Enron e-mail dataset contains 517,431 e-mails collected over three-and-a-half years from about 150 users, mostly senior management. The e-mails were all plain texts without attachments. Topics involved in the corpus included

business communication between employees, personal chats between families, technical reports, etc.

We selected the e-mails that were included in the folders named "sent", "sent_items" and "_sent_e-mail" within each user's folder. Since all users in the e-mail corpus were employees of Enron, the gender of the sender of each e-mail was validated by the name. The body of each e-mail was then parsed by removing the header, reply texts (if present) and signatures. All duplicated or carbon-copied e-mails were removed. Considering the fact that ultra-short e-mails may lack enough information and the few e-mails are lengthy, the dataset was subsequently reduced to 8970 e-mails (from 108 authors) to ensure only e-mails with more than 50 words and less than 1000 words were used for our analysis. Table 2 describes the processed e-mail corpus used in this paper.

2.3. Feature set selection

What are good linguistic features that indicate gender? This is an open research problem. Based on human psychology research (as discussed previously) and extensive experimentation, we classified five sets of gender-related features: (1) character-based; (2) word-based; (3) syntactic; (4) structure-based; and (5) function words. Table 3 tab-features lists all the 545 features we designed.

Character-based features include 29 stylometric features widely adopted in authorship attribution problems (Merriam, 1996; Tweedie et al., 1996; Vel et al., 2002), such as number of white-space characters, number of special characters (e.g., %, &.), etc.

Word-based features include 33 statistical metrics such as vocabulary richness, Yule's K measure (See Appendix A for details) and entropy measure (Baayen et al., 2002), as well as 68 psycho-linguistic features extracted from Linguistic Inquiry and Word Count (LIWC) (Online, 2007). During the last four decades, researchers have provided evidence to suggest that people's physical and mental health are correlated with the words they use (Gottschalk and Gleser, 1969; Rosenberg and Tucker, 1978). Text analysis based on these studies indicated that those individuals who benefit the most from writing tend to use relatively high rates of positive emotional words (such as love, nice, sweet), a moderate number of negative emotional words (like hurt, ugly, nasty), and an increasing number of cognitive words (like cause, know), and switch their use of pronouns from one session to another (Pennebaker et al., 2007). We calculated the feature values by counting the frequency of particular cues, and each cue may include several related words. Some examples are listed in Table 4.

Table 1 – Description of the Reuters corpus.

	Number of messages	Ave. word length per message
Male author	3474	522
Female author	3295	518
Total	6769	520

Table 2 – Description of the Enron corpus.

	Number of e-mails	Ave. word length per e-mail
Male	4947	114
Female	4023	119
Total	8970	116

Table 3 – Proposed feature sets and description.

Feature	Feature description
Character-based features	
F ₁	Total number of characters (C)
F ₂	Total number of letters(a-z)/C
F ₃	Total number of upper characters/C
F ₄	Total number of digital characters/C
F ₅	Total number of white-space characters/C
F ₆	Total number of tab space characters/C
F _{7...F₂₉}	Number of special characters (%,&,etc.)/C (23 features)
Word-based features	
F ₃₀	Total number of words (N)
F ₃₁	Average length per word (in characters)
F ₃₂	Vocabulary richness (total different words/N)
F ₃₃	Words longer than 6 characters/N
F ₃₄	Total number of short words (1-3 characters)/N
F ₃₅	Hapax legomena/N
F ₃₆	Hapax dislegomena/N
F ₃₇	Yule's K measure
F ₃₈	Simpson's D measure
F ₃₉	Sichel's S measure
F ₄₀	Honore's R measure
F ₄₁	Entropy measure
F ₄₂	The number of net abbreviation/N
F _{43...F₆₂}	Word length frequency distribution/N (20 features)
F _{63...F₁₃₀}	LIWC features (68 features)
Syntactic features	
F ₁₃₁	Number of single quotes (')/C
F ₁₃₂	Number of commas (,)/C
F ₁₃₃	Number of periods (.) /C
F ₁₃₄	Number of colons (:)/C
F ₁₃₅	Number of semi-colons (;)/C
F ₁₃₆	Number of question marks (?) /C
F ₁₃₇	Number of multiple question marks (???) /C
F ₁₃₈	Number of exclamation marks (!) /C
F ₁₃₉	Number of multiple exclamation marks (!!!!) /C
F ₁₄₀	Number of ellipsis (...) /C
Structural features	
F ₁₄₁	Total number of lines
F ₁₄₂	Total number of sentences (S)
F ₁₄₃	Total number of paragraphs
F ₁₄₄	Average number of sentences per paragraph
F ₁₄₅	Average number of words per paragraph
F ₁₄₆	Average number of characters per paragraph
F ₁₄₇	Average number of words per sentence
F ₁₄₈	Number of sentences beginning with upper case/S
F ₁₄₉	Number of sentences beginning with lower case/S
F ₁₅₀	Number of blank lines/total number of lines
F ₁₅₁	Average length of non-blank line
F ₁₅₂	Absence/present of greeting words
F ₁₅₃	Absence/present of farewell words
Function words	
F _{154...F₁₅₆}	Number of article words/N (3 features)
F _{157...F₁₆₀}	Number of pro-sentence words/N (4 features)
F _{161...F₂₃₄}	Number of pronoun words/N (74 features)
	Number of auxiliary-verbs/N (47 features)
F _{282...F₃₀₃}	Number of conjunction words/N (22 features)
F _{304...F₄₁₂}	Number of interjection words/N (109 features)
F _{413...F₅₃₆}	Number of adposition words/N (124 features)
F _{537...F₅₄₅}	Number of gender-specific words/N (9 features)

Table 4 – Examples of psycho-linguistic cues.

Feature	Words included in the feature
Negations	no, not, never
Positive emotion	love, nice, sweet
Negative emotion	hurt, ugly, nasty
Anxiety	worried, fearful, nervous
Anger	hate, kill, annoyed
Sadness	Crying, grief, sad
Insight	think, know, consider
Tentative	Maybe, perhaps, guess
Certainty	Always, never
Inhibition	block, constrain, stop
Assent	agree, OK, yes

Syntactic features capture authors' writing style at the sentence level. Syntactic features include regular punctuation (such as comma, colon, etc.) and multiple question/exclamation marks (???, !!!) since it is not uncommon for writers in very informal situations to use several question marks and exclamation marks to express the attitude or mood. The discriminating power of syntactic features is derived from man and woman's different habits of using punctuation, for example, women tend to use more question marks according to (Mulac, 1998).

Structure based features represent the way an author organizes the layout of a message. People have different habits when organizing articles. These habits, such as paragraph length and use of greetings, can be strong authorial evidence of personal writing styles. This is more prominent in online documents, which have less content information but more flexible structures or richer stylistic information. We used 13 structure-related features as listed in Table 3 tab-features.

Function words (or grammatical words) are words that have little lexical meanings or have ambiguous meanings, but instead serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker. We set function words (as listed in Appendix B) as one particular subset apart from word-based features, because function words play an important role in distinguishing the personal style of different genders. We also introduced 9 gender-linked features (Cheng et al., 2009) in this subset. For example, women make frequent use of emotionally intensive adverbs and affective adjectives such as really, very, quite and adorable, charming, lovely (Jaffe et al.). On the other hand, men's conversational patterns usually express "independence" and assertions of vertically hierarchical power, so they use more first-person singular pronouns like I and more directive sentences (Mulac et al., 1990). Some examples of gender-linked cues are listed in Table 5.

2.4. Automatic feature extraction and representation

For each message, the feature extractor produced a 545-dimension vector to represent the values of the 545 features. Since feature sets include information measured by various methods, the feature values we computed could range from 0 to more than 1000. For example, the first five feature values extracted from a message written by a male author could be 1420 0.988 0.0655 0.0119 0.217, which represents the total

Table 5 – Examples of gender-linked cues.

Feature	Words included in the feature
Affective adjectives	adorable, charming, sweet, lovely, divine
Exclamation	good heavens, hey, oh
Expletives	wow, woah
Hedges	well, kind of, sort of, possibly, maybe
Intensive adverbs	really, very, quite, special
Judgmental adjectives	distracting, bothersome, nice
Uncertainty verbs	wonder, consider, suppose

number of characters (C), the ratio of letters (total number of letters divided by C), the ratio of upper case characters (total number of upper characters divided by C), the ratio of digital characters (total number of digital characters divided by C) and the ratio of white spaces (total number of white-space characters divided by C), respectively.

To ensure all features are treated equally in the classification process, we normalized the features using max-min normalization method to ensure all feature values are between 0 and 1:

$$\text{Normalized } - x_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (2)$$

where x_{ij} is the j th feature in the i th example, $\min(x_j)$ and $\max(x_j)$ are the minimum and maximum feature values of the j th feature, separately.

2.5. Classification techniques

We designed three classifiers which are widely used in large dimension classification problem: the Bayesian logistic regression, AdaBoost decision tree and support vector machine.

2.5.1. Bayesian-based logistic regression

Logistic regression models produce an estimate of the probability that a vector x_i belongs to the class y_i :

$$P(y_i = +1 | \omega, x_i) = \psi(\omega^T x_i) \quad (3)$$

where the logistic link function is given by

$$\psi(r) = \frac{1}{1 + \exp(-r)}. \quad (4)$$

The decision of class assignment can be based on comparing the probability estimate with a threshold, i.e., predict $y = +1$ when $p(y = +1 | \omega, x_i) > \text{Threshold}$, otherwise, predict $y = -1$. We set the threshold to be 0.5 in our experiments.

One Bayesian approach to avoiding overfitting involves a prior distribution on ω that favors sparseness in the fitted model along with an optimization algorithm and implementation tailored to that prior (Genkin et al., 2007). To produce a prior favoring sparse solutions, we assume that ω_j arises from a Gaussian distribution with mean 0 and variance τ_j ,

$$p(\omega_j | \tau_j) = N(0, \tau_j), j = 1, 2, \dots, d. \quad (5)$$

Further assume that the prior τ_j 's arise from an exponential distribution with density

$$p(\tau_j | \gamma) = \frac{\gamma_j}{2} \exp\left(-\frac{\gamma_j}{2} \tau_j\right), \gamma > 0 \quad (6)$$

Integrating out τ_j then gives an equivalent nonhierarchical double-exponential (Laplace) distribution with density

$$p(\omega_j | \lambda_j) = \frac{\lambda_j}{2} \exp(-\lambda_j |\omega_j|) \quad (7)$$

where $\lambda_j = \sqrt{2}/\sqrt{\tau_j}$, and we set the prior τ_j to be the variance of sample features.

We assume that the components of ω are independent and hence the overall prior for ω is the product of the priors for each of its component ω_j 's, i.e.,

$$p(\omega) = \prod_{j=1}^d p(\omega_j | \lambda_j) = \prod_{j=1}^d \frac{\lambda_j}{2} \exp(-\lambda_j |\omega_j|) \quad (8)$$

The posterior density for ω with the logistic link on dataset \mathcal{D} is

$$\begin{aligned} L(\omega) &= p(\omega | \mathcal{D}) \propto p(\mathcal{D} | \omega) p(\omega) \\ &= \left(\prod_{i=1}^n \frac{1}{1 + \exp(-\omega^T x_i y_i)} \right) \left(\prod_{j=1}^d \frac{\lambda_j}{2} \exp(-\lambda_j |\omega_j|) \right) \end{aligned} \quad (9)$$

We get the log posterior by ignoring the normalizing constant

$$l(\omega) = -\sum_{i=1}^n \ln(1 + \exp(-\omega^T x_i y_i)) - \sum_{j=1}^d (\ln 2 - \ln \lambda_j + \lambda_j |\omega_j|) \quad (10)$$

Then ω can be estimated by finding the maximum posterior $l(\omega)$ or minimum $-l(\omega)$. Since $-l(\omega)$ is convex, a wide variety of convex optimization methods are applicable. We based our implementation on the CLG algorithm (Genkin et al., 2007), which is a one-dimensional optimization algorithm. We update each ω_j by holding all other $\omega_k, j \neq k$ constant, and traverse all ω_j 's in one pass. Multiple passes are made over ω until convergence.

2.5.2. Decision tree

Decision tree is a flowchart-like tree structure and is built by examining a measure related to information gain. In a decision tree, each attribute (or feature) is represented as an internal node, the outcome of each test is represented as a branch, and the class label is represented as a terminal node. Given a set of attribute values, a tree path is traced from the root to a terminal node that results class prediction. In general, decision tree is a well-known method for classification, and have been used in many application areas (Safavian and Landgrebe, 1991). However, the existence of high variance in the data may causes overfitting. The ensemble learning technique is induced in order to improve the classification accuracy (Damerou and Weiss, 1998).

The AdaBoost algorithm (Freund and Schapire, 1995) — from “adaptive boosting” — is one of the most important ensemble methods, since it has solid theoretical foundation, very accurate prediction, great simplicity, and wide and successful applications. Given a weak learning algorithm and a training set \mathcal{D} , the AdaBoost algorithm works as follows. First, it assigns equal weights to all the training examples $\{(x_i, y_i), (i = 1, \dots, n)\}$. Let D_t denote the distribution of the weights at the t th learning round. From the training set and D_t the algorithm generates a weak

learner $h_t: X \rightarrow Y$ by calling the decision tree algorithm. Then it uses the training examples to test h_t , and the weights of the incorrectly classified examples will be increased. Thus, an updated weight distribution D_{t+1} is obtained. From the training set and D_{t+1} , AdaBoost generates another weak learner by calling the decision tree again. Such a process is repeated for T rounds, and the final model is derived by weighted majority voting of the T weak learners, where the weights of the learners are determined during the training process.

2.5.3. Support vector machine

Support vector machine (SVM) (Cortes and Vapnik, 1995; Boser

$$\text{accuracy} = \frac{\text{number of messages whose author gender was correctly identified}}{\text{total number of messages}} \quad (14)$$

et al., 1992) is based on the structural risk minimization principle from computational learning theory. It is capable of handling large dimension inputs in linear classification problem as well as non-linear cases. As a linear classifier, if the input two classes are linearly separable, SVM maximizes the margin between the two classes by searching a linear optimal separating hyperplane. We can find an optimal weight vector ω^* by solving the following optimization problem:

$$\text{Minimize } J(\omega) = \frac{1}{2} \|\omega\|^2 \quad (11)$$

$$\text{subject to } y_i(\omega \cdot x_i - b) \geq 1$$

For the linearly nonseparable case SVM builds a soft margin (Cortes and Vapnik, 1995) by introducing the slack variable ξ and allows training examples to exist in the region between the two hyperplanes that go through the support points of the two classes. The objective function is then increased by a function which penalizes non-zero ξ_i and the optimization becomes a trade-off between a large margin, and a small error penalty. If the penalty function is linear, the problem now transforms to

$$\text{Minimize } J'(\omega) = \frac{1}{2} \|\omega^*\|^2 + C \sum_{i=1}^N \xi_i \quad (12)$$

$$\text{subject to } y_i(\omega^* \cdot x_i - b) \leq 1 - \xi, \xi \geq 0$$

When SVM is used to classify non-linear problems, the kernel trick (Aizerman et al., 1964) helps to map the feature space x into a higher-dimension space $\phi(x)$, and then SVM searches the maximum-margin hyperplane in the new space. We explored several different kernel functions, namely, linear, polynomial and radial basis functions, and obtained best results with radial basis kernel function, where all dot product $(\omega \cdot x_i)$ is replaced by the Kernel function

$$k(\omega, x_i) = \exp\left(-\frac{\|\omega - x_i\|^2}{2\delta^2}\right) \quad (13)$$

3. Experimental results

We conducted several experiments to examine the performance of different classification techniques, the impact of different combinations of parameters, and the significance of the proposed feature set. Feature extraction was implemented in Python and the classifiers were implemented in MATLAB. Each experiment was conducted 10 times by 10-fold cross validation. Through extensive experiments we fixed $\text{Gamma} = 0.5$ and $C = 3$. We defined accuracy as the measure of prediction:

3.1. Comparison of classification techniques

First, we applied the Bayesian-based logistic regression, AdaBoost decision tree and SVM classifiers separately, using Reuters and Enron corpora. The results shown in Fig. 2 indicate that SVM outperforms the other two methods for both datasets. The figure also reveals that the performance of the Bayesian logistic regression does not change significantly with the size of the training examples compared to the sharp improvement in AdaBoost (200 iterations) decision tree. The best classification result produced by the same classifier (SVM) was the accuracy of 76.75% and 82.23% respectively for two different datasets, which implies the fact that the genders of the authors of neutral news (Reuters Newsgroup) are more difficult to discriminate than the personal e-mails (Enron Corpus).

3.2. Impact of parameters

In order to study the impact of the parameters, such as the number of words per message, on the classification performance, the Enron Email corpus was further divided into 3 sub-datasets, in which each message has more than 50 words, more than 100 words, and more than 200 words, respectively. Table 6 shows the statistics measured in terms of the number of messages in each gender as a function of the minimum number of words per message.

Several experiments were conducted to evaluate the impact of the minimum number of words per message and the number of messages in the sample set on the classification performance. Generally, as the results showed, the accuracy increases as the number of words per message increases, since more words in one message may contain more information about the personal writing style and the corresponding gender influence. From Fig. 3, we can see that when the number of words per message is relatively small, i.e., when the first sub-dataset was used, the classifiers (especially SVM) can still produce moderate accuracy with small training-set size and high accuracy with large training-set. This is a good

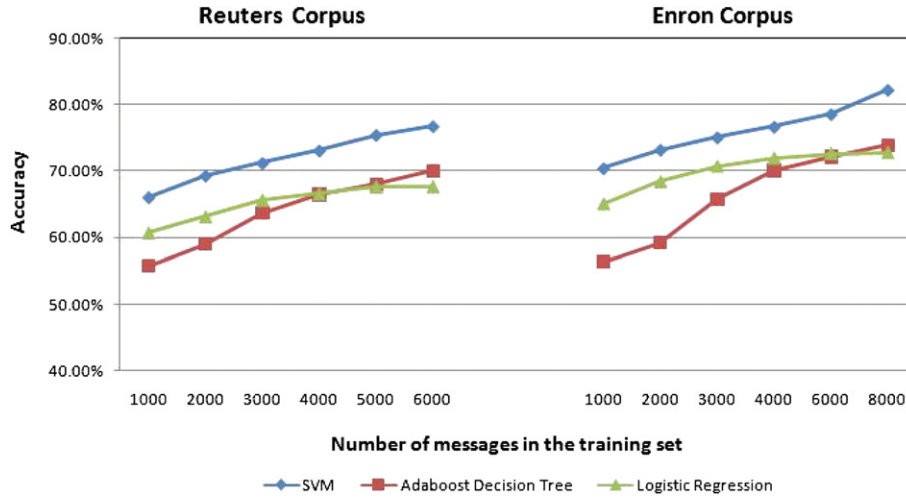


Fig. 2 – Accuracy comparison of different classifiers.

Table 6 – Statistics in terms of the number of messages per gender as a function of the minimum number of words per message.

Number of Words per E-mail	Number of E-mails		
	Male	Female	Total
Sub-dataset I <50	4947	4023	8970
Sub-dataset II <100	1978	1626	3604
Sub-dataset III <200	517	456	973

Table 7 – Accuracy comparison by using one feature subset at a time.

Feature Subset	Accuracy (%)
Word based features	59.08
Character based features	73.48
Syntactic features	65.37
Structural features	61.26
Function words	74.81
All features	85.13

indicator that the proposed model can be used to detect the gender of the author of online short messages (such as chats).

3.3. Significance of feature sets

In order to investigate the significance of the proposed feature sets, we applied SVM to the sub-dataset whose messages have a minimum of 100 words, by using one feature set at a time. The accuracies of classification are shown in Table 7. We see that all five subsets contribute to the gender identifier. The set

of word-based features and function words are shown to be important gender discriminators.

To further determine the significance of proposed features, we applied the two sample t-test to the same sub-dataset as above. By setting the significance level at 5%, we get 157 features out of 545.

When we applied feature dimension reduction by using only the significant features it resulted in faster extraction taking 1.35 seconds compared to 3.77 seconds before feature reduction. SVM classification, when applied to the sub-dataset

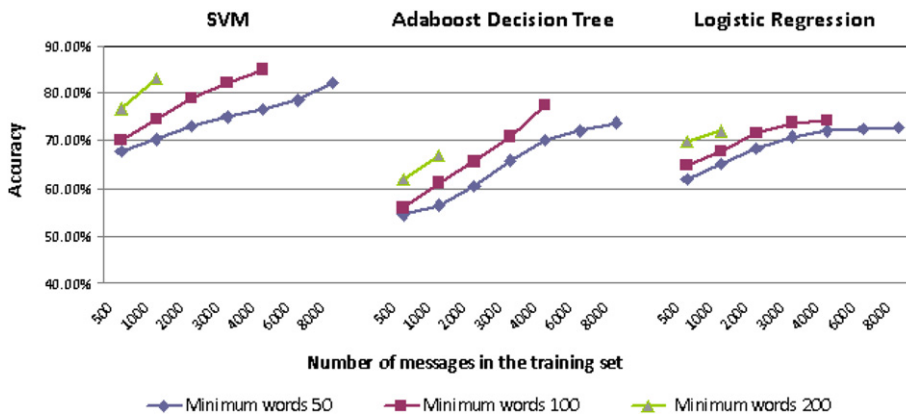


Fig. 3 – Classifier accuracy as the function of the number of words per e-mail & number of samples.

by using only the 157 significant features, resulted in an accuracy of 82.1%, a drop of 3.03% compared to the accuracy before feature reduction. Therefore we can set the cue significance level test to be higher or lower, to get the trade-off between time-cost and accuracy.

4. Conclusions

We recognize that the problem of gender identification from text is an interplay between psycho-linguistics, generic writing styles of men and women, etc. Experimental results show that SVM outperforms the Bayesian-based logistic regression and AdaBoost decision tree for identifying author’s gender from a given text document. By designing appropriate psycho-linguistic and gender-linked features, we observe that word-based features, function words and structural features play important roles in gender identification. Experimental results indicate that the identification performance is improved by increasing the number of text documents in the training dataset as well as the number of words in each document (e-mail). We find that there are significant differences between men and women in personal writings such as e-mails, and gender differences also exist between authors of news articles even though neutral language is dominant there.

Appendix A.

Some vocabulary richness measures

Yule’s K measure:

$$\text{Yules K} = 10^4 \left(-\frac{1}{N} + \sum_{i=1}^V V_i \left(\frac{i}{N} \right)^2 \right) \tag{A1}$$

Simpson’s D measure:

$$\text{Simpsons D} = \sum_{i=1}^V V_i \frac{i}{N} \frac{i-1}{N-1} \tag{A2}$$

Sichel’s S measure:

$$\text{Sichels S} = \frac{\text{count of Hapax Dislegomena}}{V} \tag{A3}$$

Honore’s R measure:

$$\text{Honores R} = \frac{100 \log_{10} N}{1 - \frac{\text{count of Hapax Legomena}}{V}} \tag{A4}$$

Entropy measure:

$$\text{Entropy} = \sum_{i=1}^N V_i \left(-\log_{10} \frac{i}{N} \right) \frac{i}{N} \tag{A5}$$

- V: number of different words
- V_i: number of different words that occur i times.
- N: total number of words
- Hapax Dislegomena: words that occur only twice
- Hapax Legomena: words that occur only once

Appendix B.

Function words

Article Words:

a	an	the
---	----	-----

Pro-sentence words:

yes	no	okay	OK
-----	----	------	----

Pronoun words:

all	everybody	his	most	other	that	what	your
another	everyone	I	much	others	theirs	whatever	yours
any	everything	it	myself	ours	them	which	yourself
anybody	few	its	neither	ourselves	themselves	whichever	yourselves
anyone	he	itself	no one	several	these	who	
anything	her	little	nobody	she	they	whoever	
both	hers	many	none	some	this	whom	
each	herself	me	nothing	somebody	those	whomever	
each other	him	mine	one	someone	us	whose	
either	himself	more	one another	something	we	you	

Auxiliary-verbs:

are	can	didn’t	hadn’t	haven’t	might	shouldn’t	won’t
aren’t	cannot	do	’d	’ve	mightn’t	was	’ll
ain’t	can’t	don’t	has	is	mustn’t	wasn’t	would
’re	could	does	hasn’t	isn’t	shall	were	wouldn’t
be	couldn’t	doesn’t	’s	’s	shan’t	weren’t	’d
been	did	had	have	may	should	will	

Conjunction words:

and	or	though	now that	if	while	in order that	in case
because	yet	unless	even though	now that	whereas	even if	
nor	so	when	although	only if	whether or not	until	

Interjection words:

adios	bah	dear	Ha-ha	howdy	oops	tush	whoosh
ah	begorra	doh	hail	hoy	ouch	tut	wow
aha	behold	duh	hallelujah	huh	phew	Tut–tut	yay
ahem	bejesus	eh	heigh-ho	humph	phooey	ugh	yikes
ahoy	bingo	encore	hello	hurray	pip–pip	uh-huh	yippee
alack	bleep	eureka	hem	hush	pooh	uh-oh	yo
alas	boo	fie	hey	indeed	pshaw	uh–uh	yoicks
all hail	bravo	gee	hey presto	jeepers creepers	rats	viva	yoo-hoo
alleluia	bye	gee whiz	hi	jeez	righto	voila	yuk
aloha	cheerio	gesundheit	hip	lo and behold	scat	wahoo	yummy
amen	cheers	goodness	hmm	man	shoo	well	zap
attaboy	ciao	gosh	ho	my word	shoot	whoa	
aw	crikey	great	ho hum	now	so long	whoopee	
ay	cripes	hah	hot dog	ooh	Touch	whoops	

Adposition words:

aboard	astride	down	of	through	worth	on to	in front of
about	at	during	off	throughout	according to	onto	in lieu of
above	athwart	except	on	till	ahead to	out from	in place of
absent	atop	failing	onto	to	as to	out of	in spite of
across	barring	following	opposite	toward	aside from	outside of	on account of
after	before	for	out	towards	because of	owing to	on behalf of
against	behind	from	outside	under	close to	prior to	on top of
along	below	in	over	underneath	due to	pursuant to	versus
alongside	beneath	inside	past	unlike	except for	regardless of	concerning
amid	beside	into	per	until	far from	subsequent to	considering
amidst	besides	like	plus	up	in to	as far as	regarding
among	between	mid	regarding	upon	into	as well as	apart from
amongst	beyond	minus	round	via	inside of	by means of	
around	but	near	save	with	instead of	in accordance with	
as	by	next	since	within	near to	in addition to	
aslant	despite	notwithstanding	than	without	next to	in case of	

REFERENCES

- Abbasi A, Chen H. Visualizing authorship for identification. IEEE International Conference on Intelligence and Security Informatics; May 2006:60–71.
- Aizerman A, Braverman EM, Rozoner LI. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control* 1964;25:821–37.
- Apte C, Damerau F, Weiss SM, Apte C, Damerau F, Weiss S. Text mining with decision trees and decision rules. In: Proceedings of the conference on automated learning and discovery, workshop 6: learning from text and the web; 1998. [Online]. Available: http://www.50x15.com/en-us/internet_usage.aspx.
- Baayen H, van Halteren H, Neijt A, Tweedie F. An experiment in authorship attribution. In: Proceedings of the 6th international conference on the statistical analysis of textual data; 2002.

- Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: Proceedings of the 5th annual ACM workshop on computational learning theory. ACM Press; 1992. p. 144–52.
- Burrows J. Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing* 1987;2:61–7.
- Chen H. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems: Special Issue on AI for Homeland Security* 2005;5(20):67–75.
- Cheng N, Cheng X, Chandramouli R, Subbalakshmi K.P. “Gender identification from e-mails,” in *IEEE Symposium on computational intelligence and data mining proceedings*, 2009, pp. 154–158.
- Cortes C, Vapnik V. Support-vector networks. In: *Machine learning*; 1995. p. 273–97.
- Crawford M. *Talking difference: on gender and language*. London: Sage; 1995.
- Damerau Apte F, Weiss S. *Text mining with decision trees and decision rules*; 1998.
- Diederich J, Kindermann J, Leopold E, Paass G. Authorship attribution with support vector machines. *Applied Intelligence* 2000;19:109–23.
- Efron R, Thisted B. Estimating the number of unseen species: how many words did shakespeare know? *Biometrika* 1976;63(3): 435–47.
- Freund Y, Schapire RE. *A decision-theoretic generalization of on-line learning and an application to boosting*; 1995.
- Genkin Alexander, Lewis David, Madigan D, David. Large-scale bayesian logistic regression for text categorization. *Technometrics* 2007;49(3):291–304 [Online] Available, <http://dx.doi.org/10.1198/004017007000000245>.
- Gottschalk LA, Gleser GC. *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley: University of California Press; 1969.
- Holmes D. A stylometric analysis of mormon scripture and related texts, vol. 155. *Royal Statistical Society*; 1992. pp. 91–120.
- Holmes DI, Forsyth R. The federalist revisited: new directions in authorship attribution. *Literary and Linguistic Computing* 1995;10(2):111–27.
- Jaffe M, Lee Y, Huang L, Oshagan H. “Gender, pseudonyms, and cmc: masking identities and baring souls,” *Gender, Language, & CMC*. [Online]. Available: <http://research.haifa.ac.il/jmjaffe/genderpseudocmc/gender.html>.
- Lakoff R. *Language and woman’s place*. New York: Harper and Row; 1975.
- Lowe D, Matthews R. Shakespeare vs. fletcher: a stylometric analysis by radial basis functions. *Computers and the Humanities* 1995;29:449–61.
- Mendenhall TC. The characteristic curves of composition. *Science* 1887;11(9):237–46.
- Merriam T. Marlowe’s hand in Edward III revisited. *Literary and Linguistic Computing* 1996;11(1):19–22.
- Mosteller F, Wallace DL. *Inference and disputed authorship: the federalist*. Reading, MA: Addison-Wesley Publishing Company, Inc.; 1964.
- Mosteller F, Wallace DL. *Applied bayesian and classical inference: the case of the federalist papers*, ser. In: *Springer series in statistics*. Springer; 1984.
- Mulac A. The gender-linked language effect: do language differences really make a difference?; 1998.
- Mulac A, Lundell TL. Effects of gender-linked language differences in adults’ written discourse: multivariate tests of language effects. *Language and Communication* 1994;14(3).
- Mulac A, Studley LB, Blau S. The gender-linked language effect in primary and secondary students’ impromptu essays. *Sex Roles* 1990;23(9–10).
- Newman ML, Pennebaker JW, Berry DS, Richards JM. Lying words: predicting deception from linguistic styles. *Personality and Social Psychology Bulletin* 2003;29:665C675.
- Reuters corpora [Online]. Available, <http://trec.nist.gov/data/reuters/reuters.html>; 2000.
- Bibliography of gender and language [Online]. Available, <http://ccat.sas.upenn.edu/~haroldfs/popcult/bibliogs/gender/genbib.htm>; 2002, July.
- Linguistic inquiry and word count [Online]. Available, <http://www.liwc.net/>; 2007, Jun.
- Enron e-mail dataset [Online]. Available, <http://www-2.cs.cmu.edu/~enron/>; 2005, April.
- Peng F, Schuurmans D, Keselj V, Wang S. Automated authorship attribution with character level language models. In: *Proceedings of the 10th conference of the European chapter of the association for computational linguistics*; 2003.
- Pennebaker J. *Emotion, disclosure, and health*; 1995.
- Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ. *The development and psychometric properties of LIWC2007*. Austin, Texas: LIWC Inc; 2007.
- Rosenberg SD, Tucker GJ. Verbal behavior and schizophrenia: the semantic dimension. *Archives of General Psychiatry* 1978;36: 1331–7.
- Safavian SR, Landgrebe D. A survey of decision tree classifier methodology; May 1991. no. 3660–674.
- Talbot MM. *Language and gender: an introduction*. Wiley-Blackwell; 1998.
- Tweedie FJ, Singh S, Holmes DI. Neural network applications in stylometry: the federalist papers. *Computers and the Humanities* 1996;30(1):1–10.
- Vel OD, Corney M, Anderson A, Mohay G. Language and gender author cohort analysis of e-mail for computer forensics. In: *Proc. digital forensic research workshop*; 2002. [Online]. Available: <http://www.foxnews.com/story/0,2933,312018,00.html>.
- Yule GU. *The statistical study of literary vocabulary*. Cambridge University Press; 1944.
- Zheng R, Li J, Chen H, Huang Z. A framework for authorship identification of online messages: writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology* 2006;57(3): 378–83.