

# Gender Identification from E-mails

Na Cheng, Xiaoling Chen, R. Chandramouli, K. P. Subbalakshmi  
Department of ECE  
Stevens Institute of Technology  
Hoboken, NJ 07030

**Abstract**—In this paper, we investigate the topic of gender identification for short length, multi-genre, content-free e-mails. We introduce for the first time (to our knowledge), psycholinguistic and gender-linked cues for this problem, along with traditional stylometric features. Decision tree and Support Vector Machines learning algorithms are used to identify the gender of the author of a given e-mail. The experiment results show that our approach is promising with an average accuracy of 82.2%.

## I. INTRODUCTION

THE rapid development of Internet has created myriad ways to distribute information across time and space. E-mail is one of the most widely used web-based channels for information exchange. Trillions of business letters, financial transactions, governmental orders and private messages are exchanged through e-mail system each year [1].

Unfortunately, the benefit of e-mails is being threatened by various kinds of misuses, such as e-mail bombardment, spamming, phishing and e-mail worms. The New York Times reported in April 2008 that “e-mail has become the bane of some people’s professional lives” [2].

In many misuses or crime cases, the senders use anonymous e-mail servers to hide their address information, and conceal his/her true identity (such as name, age and gender) to avoid being detected. In such a situation, it becomes imperative to design efficient automated methods to track senders’ identities within the environment of cyberspace. Authorship identification based on the analysis of stylometric features is one such approach.

The origins of authorship identification studies dates back to the 18th century when English logician Augustus de Morgan suggested that authorship might be settled by determining if one text contained more longer words than another [3]. Well known authorship identification studies include the attribution of disputed Shakespearean works (e.g. [4], [5], [6]) and the attribution of the Federalist papers ([7], [8], [9]).

Although authorship identification methods have achieved great success in many literary and forensic applications as mentioned above, very limited studies have been undertaken specifically for short messages such as e-mails.

In this paper, we are particularly interested in identifying the gender of the authors of e-mails. Machine learning algorithms are proposed to address this problem. The rest of this paper is organized as follows: Section II describes the formulation of the gender identification problem. Section III introduces the dataset we use in our experiments, and

also describes a preprocessing procedure. In Section IV we describe the selection of feature sets. Two classification methods and their implementations are discussed in Section V and Section VI respectively. Concluding remarks are provided in Section VII.

## II. PROBLEM FORMULATION

Generally, men and women converse differently even though they technically speak the same language. Many studies have been undertaken to study the relationship between gender and language use [10]. Empirical evidence suggests the existence of gender differences in written communication, face-to-face interaction and computer-mediated communication [11].

Gender identification problem can be treated as a binary classification problem in (1), i.e., given two classes  $\{male, female\}$ , assign an anonymous e-mail  $e$  to one of them according to the gender of the corresponding author:

$$e \in \begin{cases} Class_1 & \text{if the author of } e \text{ is male} \\ Class_2 & \text{if the author of } e \text{ is female} \end{cases} \quad (1)$$

To test the binary hypothesis (1), we have to select a set of features that remain relatively constant for a large number of e-mails written by authors of the same gender. Once the feature set has been chosen, a given e-mail can be represented by an  $n$ -dimensional vector, where  $n$  is the total number of features. Given a set of known pre-classified e-mails, a model (or classifier) can be built by classification techniques, which can then be used to determine the category of a new e-mail.

In general, the procedure of gender identification process can be divided into four steps:

- 1) Collect a suitable corpus of e-mails as dataset
- 2) Identify significant features in distinguishing genders
- 3) Extract feature values from each e-mail automatically
- 4) Build a classification model to identify the gender of the author of any e-mail

## III. PRE-PROCESSING ENRON E-MAIL CORPUS

The collection of a suitable corpus of e-mails is typically limited by privacy and ethical considerations. Publicly, there is no large scale e-mail dataset readily available for gender identification research. We mitigate this problem by using the Enron e-mail dataset [12].

Enron was an energy company based in Houston, Texas. Enron 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.

The e-mail corpus turned out to have a number of integrity problems, which were corrected by a number of researchers later. Some messages have been deleted “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.

Enron e-mail dataset contains 517,431 e-mails from about 150 users, mostly senior management. The e-mails are all plain texts without attachments. Topics involved in the corpus include business communication between employees, personal chats between families, technical reports, etc.

We select the e-mails that are included in the folders called “*sent*”, “*sent\_items*” and “*\_sent\_email*” within each user’s folder in our experiment. Since all users in the e-mail corpus were employees of Enron, the gender of the sender of each e-mail can be 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 length of e-mails are commonly not long, the dataset was subsequently reduced to 8,970 e-mails (from 108 authors) to ensure only e-mails with more than 50 words and less than 1000 words are used for our analysis. Table I describes the e-mail corpus used in our experiment.

TABLE I  
DESCRIPTION OF THE E-MAIL CORPUS

	<i>Number of Emails</i>	<i>Number of Authors</i>	<i>Ave. Number of Emails per Author</i>	<i>Ave. word length per Email</i>
<i>Male</i>	4947	86	58	114
<i>Female</i>	4023	28	144	119
<i>Total</i>	8970	114	79	116

In order to study the impact of the number of words in an e-mail on the classification performance, the e-mail corpus was further divided into multiple sub-datasets, of which, each e-mail has more than 50 words, more than 100 words, and more than 200 words separately. Table II shows the statistics measured in terms of the number of e-mails in each gender as a function of the minimum number of words per e-mail.

#### IV. FEATURE SET SELECTION AND ANALYSIS

It has been suggested by previous research that distinguishing characteristics of male/female linguistic styles exist [10]. Gender-linked effects on language were introduced in [13], [14] by analyzing students’ impromptu essays. Both fiction and nonfiction corpora (British National Corpus) were evaluated in [15], and the results indicated that the use of

TABLE II  
STATISTICS IN TERMS OF THE NUMBER OF E-MAILS IN EACH GENDER AS A FUNCTION OF THE MINIMUM NUMBER OF WORDS PER E-MAIL

<i>Number of Words per E-mail</i>	<i>Number of E-mails</i>		
	<i>Male</i>	<i>Female</i>	<i>Total</i>
<i>sub-dataset I &gt; 50</i>	4947	4023	8970
<i>sub-dataset II &gt; 100</i>	1978	1626	3604
<i>sub-dataset III &gt; 200</i>	517	456	973

pronouns and certain types of noun modifiers are different between male and female authors.

To identify the gender of the author of an e-mail is different from the other types of authorship identification problems. First, the length of e-mail is usually very short compared to other types of texts like books and novels. Second, the style of e-mails may change a lot according to the type or social status of recipients, for example, formal style in business e-mails and informal style in personal e-mails. Third, some special linguistic elements such as facial expressions often appear in e-mails. Fourth, the format or the structure of e-mails may vary among different users. Thus, specific email-based gender-differentiating feature sets must be considered along with traditional stylometric features.

The traditional stylometric features are described in detail in [3]. We divide the features into five subsets: character based features, word based features, syntactic based features, structure based features and function words.

In word based feature-set, psycho-linguistic features are introduced in this paper for the first time for the gender identification problem. 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 [16], [17]. Text analysis based on these studies indicate that those individuals who benefit the most from writing tend to use relatively high rates of positive emotion words (such as *Love, nice, sweet*), a moderate number of negative emotion 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 session [18]. In our experiment we extract 68 psycho-linguistic features using a text analysis tool, called Linguistic Inquiry and Word Count (LIWC) [19]. Each feature may include several related words, and some examples are listed in Table III.

TABLE III  
EXAMPLES OF LIWC FEATURES

<i>Feature</i>	<i>words included in the feature</i>
Negations	no, not, never
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

TABLE IV  
FEATURE SETS AND DESCRIPTION

Feature	Feature description
<u>Character based features</u>	
$F_1$	total number of characters in words(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
$F_7 \dots F_{29}$	number of special characters(% ,etc.)/C(23 features)
<u>Word based features</u>	
$F_{30}$	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
$F_{43} \dots F_{62}$	word length frequency distribution/N (20 features)
$F_{63} \dots F_{130}$	LIWC features (68 features)
<u>Syntactic features</u>	
$F_{131}$	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
<u>Structural features</u>	
$F_{141}$	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
<u>Function words</u>	
$F_{154} \dots F_{156}$	number of article words/N (3 features)
$F_{157} \dots F_{160}$	number of pro-sentence words/N (4 features)
$F_{161} \dots F_{234}$	number of pronoun words/N (74 features)
$F_{235} \dots F_{281}$	number of auxiliary-verbs/N (47 features)
$F_{282} \dots F_{303}$	number of conjunction words/N (22 features)
$F_{304} \dots F_{412}$	number of interjection words/N (109 features)
$F_{413} \dots F_{536}$	number of adposition words/N (124 features)
$F_{537} \dots F_{545}$	number of gender-specific words/N (9 features)

Besides, we also introduce 9 gender-linked features [20] as part of the function words feature-set. For example, the female make frequent use of emotionally intensive adverbs and affective adjectives such as *really*, *very*, *quite* and *adorable*,

*charming*, *lovely*. On the other hand, male 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.

Table IV lists all the 545 features we used in our experiment.

## V. GENDER CLASSIFICATION METHODS

The classification models employed in previous contributions on authorship attribution can be divided into two broad categories: statistical methods and machine learning methods. Statistical methods were widely used in early studies based on histograms of word-length distribution of various authors [21]. Other statistical approaches applied in this field include Bayesian classifier [22], Principle Component Analysis [23], cluster analysis [24], etc. The advent of powerful computers instigated the extensive use of machine learning techniques in authorship analysis, such as neural network [9], Support Vector Machine (SVM) [25], etc. In general, machine learning methods can deal with a larger set of features with fewer requirements on mathematical models or assumptions [26] and are tolerant to noise and nonlinear interactions among features [27].

In our experiment, we tried two popular machine learning algorithms: decision tree and SVM. 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 classifiers can handle high dimensional dataset, and thus have been used in many application areas. However, decision tree is still a weak learner. In order to improve the classification accuracy, we use ensemble classifiers by employing adaptive boosting, where the training set is selected based on the error of the previous trained hypothesis, and higher weights are given to “difficult” examples.

The other machine learning algorithm we applied is SVM, which is a strong learner for both linear and nonlinear data classification. When the input attributes of two classes are linearly separable, SVM maximizes the margin between the two classes by searching a linear optimal separating hyperplane. On the other hand, when the input attributes of two classes are linearly inseparable, SVM will first map the feature space into a higher-dimension space by a nonlinear mapping, and then search the maximum-margin hyperplane in the new space. By choosing an appropriate nonlinear mapping function, input attributes from two classes can always be separated. In the gender identification problem, we explored several different kernel functions, namely, linear, polynomial and radial basis functions, and obtained best results with radial basis kernel function, which is defined as:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{2\delta^2}\right) \quad (2)$$

## VI. EXPERIMENT DESIGN AND SIMULATION RESULTS

We examine the performance of different classification techniques, the impact of the number of words in each e-mail, the number of e-mails in the sample set, and the effect of the choice of feature sets for gender identification accuracy. Each experiment is conducted 10 times by 10-fold cross validation.

First, we apply decision tree and SVM classifiers separately, by using all the e-mails in the three sub-datasets presented in Section III. The results shown in Table V indicate that SVM outperforms decision tree for all the sub-datasets.

TABLE V  
ACCURACY (%) COMPARISON BETWEEN DECISION TREES AND SVM

Classifier	minimum words per e-mail		
	50	100	200
Decision Tree	73.38	80.43	78.93
SVM	80.08	82.20	81.03

Several experiments were conducted to evaluate the impact of the minimum number of words per e-mail and the number of e-mails in the sample set on the classification performance. Generally, the results show that the accuracy increases as the number of words per e-mail increases, or the number of e-mails in sample set increases. From Fig.1, we can see that when the number of words per e-mail is relatively small, the performance of gender identification is still satisfactory.

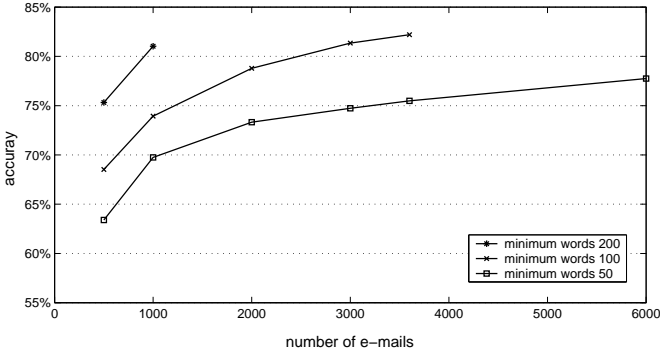


Fig. 1. Classifier accuracy as the function of the number of words per e-mail & number of samples

In order to investigate the significance of each one of the proposed feature sets, we applied SVM to the sub-dataset whose e-mails have a minimum of 100 words, by removing one feature set at a time. The classification accuracies are shown in Table VI. We see that the removal of any of the feature reduces the performance. The set of function words and word-based features are seen to be important gender discriminators.

TABLE VI  
ACCURACY COMPARISON BY REMOVING ONE FEATURE SET

Feature Set Removed	Accuracy(%)	Difference(%)
None	82.20	0
Function words	78.79	-3.41
Word based features	78.98	-3.22
structural features	79.28	-2.92
syntactic features	80.94	-1.26
Character based features	81.63	-0.57

To evaluate the overall performance over all binary categorization tasks, we use the macro-averaged  $F_1$  measure [28], i.e.,

$$\overline{F_1}^M = \frac{\sum_{i=1}^{N_{GC}} F_{1,GC_i}}{N_{GC}} \quad (3)$$

Where  $N_{GC}$  is the number of gender categories (GC). In our case,  $N_{GC} = 2$  and  $GC_1 = male$ ,  $GC_2 = female$ .  $F_{1,GC_i}$  is defined as

$$F_{1,GC_i} = \frac{2R_{GC_i}P_{GC_i}}{R_{GC_i} + P_{GC_i}} \quad (4)$$

Where  $R_{GC_i}$  and  $P_{GC_i}$  denote the Precision (P) and Recall (R) of the  $i$ th category. E.g., for  $GC_1$ , a confusion matrix can be constructed as Table VII, the numbers along the major diagonal represent the correct decisions made, and the numbers off this diagonal represent the errors-the confusion-between the various classes. Then precision and recall for this category can be calculated from the matrix:

$$P_{GC_1} = \frac{TP}{TP + FP} \quad (5)$$

$$R_{GC_1} = \frac{TP}{TP + FN} \quad (6)$$

TABLE VII  
CONFUSION MATRIX

		Predict Class	
		Male	Female
Actual Class	Male	True Positive(TP)	False Negative(FN)
	Female	False Positive(FP)	True Negative(TN)

As observed in Table VIII,  $F_1$  for male class is slightly better than the female class, which indicates the ability for identifying male authors is slightly better than identifying female authors. The table also shows that measurements by  $\overline{F_1}^M$  and accuracy are consistent.

TABLE VIII  
COMPARISON AMONG DIFFERENT MEASURE METRICS

$F_{1,male}$	$F_{1,female}$	$\overline{F_1}^M$	accuracy
83.91%	78.99%	81.45%	82.20%

## VII. CONCLUSIONS

In this paper, we investigated the gender identification problem by using the Enron e-mail dataset. Results show that SVM outperforms the decision tree method. By introducing psycho-linguistic and gender-linked features, we observed that word-based features and function words play important roles in gender identification. Experiment results indicate that the identification performance is improved by increasing the number of e-mails in the training data set as well as the number of words in each e-mail.

A lot of work still remains and needs to be done in the future to improve the accuracy of gender identification. We plan to explore additional features such as cues exhibited by the writing pattern of sentences.

## ACKNOWLEDGEMENT

This work was supported by the Ravenshield Project.

## REFERENCES

- [1] G. U. YULE, "A novel approach of mining write-prints for authorship attribution in e-mail forensics," *Digital Investigation*, vol. 5, pp. 42–51, 2008.
- [2] S. Goodson, "Email has become the bane of people's professional lives," *New York Times*, April 2008.
- [3] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.
- [4] R. Efron and B. Thisted, "Estimating the number of unseen species: How many words did shakespeare know?" *Biometrika*, vol. 63, no. 3, pp. 435–447, 1976.
- [5] D. Lowe and R. Matthews, "Shakespeare vs. fletcher: A stylometric analysis by radial basis functions," *Computers and the Humanities*, vol. 29, pp. 449–461, 1995.
- [6] T. Merriam, "Marlowe's hand in Edward III revisited," *Literary and Linguistic Computing*, vol. 11, no. 1, pp. 19–22, 1996.
- [7] F. Mosteller and D. L. Wallace, *Inference and Disputed Authorship: The Federalist*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1964.
- [8] D. I. Holmes and R. Forsyth, "The federalist revisited: New directions in authorship attribution," *Literary and Linguistic Computing*, vol. 10, no. 2, pp. 111–127, 1995.
- [9] F.J.Tweedie, S.Singh, and D.I.Holmes, "Neural network applications in stylometry: The federalist papers," *Computers and the Humanities*, vol. 30, no. 1, pp. 1–10, 1996.
- [10] (2002, July) Bibliography of gender and language. [Online]. Available: <http://ccat.sas.upenn.edu/~haroldfs/popcult/bibliogs/gender/genbib.htm>
- [11] M. Corney, O. Vel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *18th Annual Computer Security Applications Conference*, 2002, pp. 21–27.
- [12] (2005, April) Enron email dataset. [Online]. Available: <http://www-2.cs.cmu.edu/~enron/>
- [13] A. Mulac, L. B. Studley, and S. Blau, "The gender-linked language effect in primary and secondary students' impromptu essays," *Sex Roles*, vol. 23, no. 9-10, 1990.
- [14] A. Mulac and T. L. Lundell, "Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects," *Language and Communication*, vol. 14, no. 3, 1994.
- [15] S. Argamon and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.
- [16] L.A.Gottschalk and G.C.Gleser, *The measurement of psychological states through the content analysis of verbal behavior*. Berkeley: University of California Press, 1969.
- [17] S.D.Rosenberg and G.J.Tucker, "Verbal behavior and schizophrenia: The semantic dimension," *Archives of General Psychiatry*, vol. 36, pp. 1331–1337, 1978.
- [18] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, *The Development and Psychometric Properties of LIWC2007*. LIWC Inc, Austin, Texas, 2007.
- [19] (2007, Jun) Linguistic inquiry and word count. [Online]. Available: <http://www.liwc.net/>
- [20] A.Mulac, J.Bradac, and P.Gibbons, "Empirical support for the gender as culture hypothesis: An intercultural analysis of male/female language differences," *Human Communication Research*, vol. 27, pp. 121–152, 2001.
- [21] M. TC, "The characteristic curves of composition," *Science*, vol. 11, no. 9, pp. 237–246, 1887.
- [22] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, ser. Springer Series in Statistics. Springer, 1984.
- [23] J. Burrows, "Word patterns and story shapes: The statistical analysis of narrative style," *Literary and Linguistic Computing*, vol. 2, pp. 61–67, 1987.
- [24] D. Holmes, "A stylometric analysis of mormon scripture and related texts," *Royal Statistical Society*, vol. 155, pp. 91–120, 1992.
- [25] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Applied Intelligence*, vol. 19, pp. 109–123, 2000.
- [26] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.
- [27] D. L. MEALAND, "Correspondence analysis of luke," *Literary & Linguistic Computing*, vol. 10, pp. 171–182, 1995.
- [28] Y. Yang, "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, vol. 1, pp. 67–88, 1999.