International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359

Research Article | **January 2013**

Appropriate Gender Identification from the Text

Dr. G. Murugaboopathy, *Head-R&D, Veltech Multitech Dr. Rangarajan DR Sakunthala Engineering College, India*
S.Hariharasitaraman, *Asst Prof-II, Dept of IT, Kalasalingam University, India*
Dr.N.Sankarram, *Prof &Head-CSE, RMD Engineering College, Chenna, India*
T.K.S.Rathish Babu, *Asst Prof, Veltech Multitech Dr R. S. Engineering College, India*

Abstract-

This paper describes an investigation of authorship gender attribution mining from e-mail text documents. A set of topic, content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features are used. Support Vector Machine learning algorithm is used. Experiments using a corpus of e-mail documents generated by a large number of authors of both genders performed for author gender categorization.

Keywords: Data mining, Feature set, Authorship Identification, SVM.

I. Introduction

Data mining is the process of extracting patterns from data. The understandable patterns are used to make predictions or classifications about new data, explain existing data, summarize the contents of a large database to support decision making, graphical data visualization to aid humans in discovering deeper patterns. One of the applications is Text mining (news group, email, documents) and Web mining Authorship attribution, the science of inferring characteristics of the author from the characteristics of documents written by that author, is a problem with a long history and a wide range of application With the rise in the use of computers and computer networks for illegal activities (e.g., fraud) the area of computer forensics has become increasingly important. Computer forensics investigations have to increasingly deal with e-mail as this is becoming an important form of communication for many computer users, [2]E-mail is used in many legitimate activities such as message and document exchange. Unfortunately, it can also be misused for example, mailing of offensive or threatening material. In many misuses the senders use anonymous e-mail servers to hide their address information, and mask his/her true identity (such as name, age, gender, social status) to avoid being detected.

In such a situation, it becomes very important to design efficient automated methods to track senders' identity within the environment. In this paper we are interested in identifying gender information from emails. Since Machine learning algorithms gave promising results for authorship attribution, [3]we proposed that to address this gender classification problem.

The paper is organized as Section II describes Problem formulation, Section III describes data set preprocessing, Section IV comprises of features sets used and section V & VI represents methods and conclusions.

II. Problem Formulation

Gender identification problem can be treated as a binary classification problem i.e., given two classes {male, female}. To test the binary suggestion, 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 selected, a given e-mail can be represented by an n-dimensional vector, where n is the total number of features.

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
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

The used 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.

The e-mails that are included in the folders called "sent", "sent items" and "sent email" within each user's folder are selected[5]. The body of each email 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 emails are commonly not long, the dataset was subsequently reduced to ensure only e-mails with more than 50 words and less than 1000 words are used for our analysis.

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.

IV. Feature Set Selection

Gender-linked effects on language were introduced. Gender identification 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 according to the type or social status of recipients, for example, formal style in business e-mails and informal style in personal emails. 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 features are divided into five subsets: character based features, word based features, syntactic based features[7], and structure based features and function words.

In word based feature-set, psycho-linguistic features are introduced. During the last four decades, researchers have provided evidence to suggest that people's physical and mental health is correlated with the words they use. 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[8] their use of pronouns from one session to another session.

Gender-linked features are used as part of the function words feature-set.

Character based features

- total number of characters in words(C)
- total number of letters(a-z)/C
- total number of upper characters/C
- total number of digital characters/C
- total number of white-space characters/C
- total number of tab space characters/C
- number of special characters(%,etc.)/C

Word based features

- total number of words (N)
- average length per word (in characters)
- vocabulary richness (total different words/N)
- words longer than 6 characters/N
- total number of short words (1-3 characters)/N
- number of net abbreviation/N

Syntactic features

- number of single quotes(') /C
- number of commas(,)/C
- number of periods(.)/C
- number of colons(:)/C
- number of semi-colons(;)/C
- number of question marks(?)/C
- number of multiple question marks(???)/C
- number of exclamation marks(!)/C
- number of multiple exclamation marks(!!!!)/C
- number of ellipsis(. . .) /C

Structural features

- total number of lines
- total number of sentences (S)
- total number of paragraphs
- average number of sentences per paragraph
- average number of words per paragraph
- average number of characters per paragraph
- average number of words per sentence
- number of sentences beginning with upper case/S
- number of sentences beginning with lower case/S

- number of blank lines/total number of lines
- average length of non-blank line

Function words

- number of article words/N
- number of pronoun words/N
- number of auxiliary-verbs/N
- number of conjunction words/N
- number of interjection words/N
- number of gender-specific words/N

LIWC features

Feature Words included

Negations no, not, never Anxiety worried, fearful, Anger hate, kill, annoyed Sadness crying, grief, sad Insight think, know, consider Tentative maybe, perhaps, guess Gender Linked Features

Male	**Female**
Key points	Full context
Headlines	Complete article
Facts & features	Stories & personal details
Independence and assertions	Emotionally intensive adverbs

V. Methods

The classification models employed can be divided into two broad categories: statistical methods and machine learning methods. In general, machine learning methods can deal with a larger set of features with fewer requirements on mathematical models or assumptions. In our approach, two popular machine learning algorithms are used: decision tree and SVM. Decision tree is a flowchart-like tree structure [9]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[6]. 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 suggestions, and higher weights are given to "difficult" examples.

The other machine learning algorithm 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 hyper plane[11]. 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 hyper plane 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.

VI. Conclusions

The gender identification problem by using SVM.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. Generally gender 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. We plan to explore additional features such as cues exhibited by the writing pattern of sentences.

REFERENCES

[1] Dlker Nadi Bozkurt, Ozgur Baglıoglu, Erkan Uyar,"Authorshipattribution" Computer and Information Sciences, pp.1-5,2007.

[2] 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.

[3] Robert Goodman, Matthew Hahn, Madhuri Marella, Christina Ojar, Sandy Westcott,The Use of Stylometry for Email Author Identification: A Feasibility Study Proceedings of Student/Faculty Research Day, CSIS, Pace University,2007
[4] Kjell, **B.**Authorship attribution of text samples using neural networks and Bayesian classifiers**;** Systems, Man, and Cybernetics, Humans, Information and Technology, vol.2, pp.1660-1664, 1994.
[5] Sanderson, C.; Guenter, S.On Authorship Attribution via Markov Chains and Sequence Kernels; Pattern Recognition, vol.3,pp.437-440,2006.
[6] 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.
[7] 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.
[8] Y. Yang, "An evaluation of statistical approaches to text categorization," Journal of Information Retrieval, vol. 1, pp. 67–88, 1999.
[9] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," Applied Intelligence, vol. 19, pp. 109–123, 2000.
[10] (2007, Jun) Linguistic inquiry and word count. Available: http://www.liwc.net/
[11] Wikipedia.org
[12] The classic vector space model.http://www.miislita.com/termvector/term-vector-3.html .
[13] Shlomo Argamon, Levitan Shlomo. Measuring the usefullness of function words for Authorship Attribution. Proceedings of ACH/ALLC Conference ,2000.
[14] TC joachims "Text categorization with Support vector machines", European conference on machine learning,1998.
[15] Bryan Klimt and Yiming Yang, The Enron Corpus:A New Dataset for Email Classification Research.