Chat Mining for Gender Prediction

Tayfun Kucukyilmaz, B. Barla Cambazoglu,
Cevdet Aykanat, and Fazli Can

Bilkent University, Department of Computer Engineering,
06800 Bilkent, Ankara, Turkey
{ktayfun, berkant, aykanat, canf}@cs.bilkent.edu.tr

Abstract. The aim of this paper is to investigate the feasibility of predicting the gender of a text document's author using linguistic evidence. For this purpose, term- and style-based classification techniques are evaluated over a large collection of chat messages. Prediction accuracies up to 84.2% are achieved, illustrating the applicability of these techniques to gender prediction. Moreover, the reverse problem is exploited, and the effect of gender on the writing style is discussed.

1 Introduction

Authorship characterization is a problem long studied in literature [1]. In general terms, authorship characterization can be defined as the problem of predicting the attributes (e.g., biological properties and socio-cultural status) of the author of a textual document. The outcome of such studies are primarily used for financial forensics, law enforcement, threat analysis, and prevention of terrorist activities. Consequently, in several studies [2,3], efforts have been spent to increase the prediction accuracies in authorship characterization.

In this paper, we investigate the problem of predicting the gender of a text document's author. In particular, we focus on the text-based communications over the Internet. This type of communications are observed in online services such as MSN messenger, ICQ, and media supporting written discourse such as email, newsgroups, discussion forums, IRCs, and chat servers. We first formulate the problem as a text classification problem, in which the words in a document are used to attribute a gender to the author of the document. Second, we investigate the effect of stylistic features (e.g., word lengths, the use of punctuation marks, and smileys) on predicting the gender. Finally, we exploit the reverse problem and discuss the effect of gender on the writing style.

The rest of the paper is organized as follows. In Section 2, we provide a short literature survey of the studies on authorship analysis. In Section 3, we present the dataset used in this work and define the gender prediction problem. The techniques we employed to solve the problem are presented in Section 4. Section 5 provides the results of a number of experiments conducted to evaluate the feasibility of gender prediction. We finalize the paper in Section 6 with a concluding discussion on the effect of gender on the writing style.

T. Yakhno and E. Neuhold (Eds.): ADVIS 2006, LNCS 4243, pp. 274–283, 2006.

2 Related Work

The authorship studies in literature can be divided into three categories [2]: authorship attribution, similarity detection, and authorship characterization. The authorship attribution is the task of finding or validating the author of a document. Some well-known examples of authorship attribution are the examination of Shakespeare's works [4,5] and the identification of the authors of the disputed Federalist papers [6,7,8]. Similarity detection aims to find the variation between the writings of an author [9] or to differentiate between the text segments written by different authors [10], mostly for the purpose of detecting plagiarism.

Authorship characterization can be defined as the task of assigning the writings of an author into a set of categories according to his sociolinguistic profile. Some attributes previously analyzed in literature are gender, educational level, language origin, and cultural background. In [11], gender and language origin of authors are examined using machine learning techniques. In [12], English text documents are classified according to the author's gender and document's genre. In [13], a set of documents are classified according to their genre under legal, fictional, scientific, and editorial categories.

Authorship studies took more attention with the widespread use of computers, which led to an explosion in the amount of digital text documents (e.g., emails, program codes, chat messages, posts on the forums). In literature, several studies addressed the analysis of these documents based on the writing styles of the authors. In [14], identities of programmers are questioned using several stylistic features such as the use of comments, selection of variable names, and use of programming constructs. In [2,11], a collection of email documents are examined for predicting the identity and gender of the author. The typical features used are message tags, signatures, and the vocabulary richness.

3 Dataset and Problem Definition

The chat dataset used in this paper is collected from a chat server (Heaven BBS), where users have peer-to-peer communication via textual messages. The dataset consists of a collection of message logs storing the users' outgoing messages (typed in Turkish). The messages are logged for a one-month period without the notice of the users, but respecting the anonymity of the users and messages. There are around 1500 users, each with a subscription information including personal details such as gender, age, and occupation. The vocabulary of the dataset contains about 50,000 distinct words, consisting of only ASCII characters. There are around 250,000 chat messages, which are usually very short (6.2 words per message on the average).

In this paper, our aim is to find a classification of users according to their gender by using both term-based and style-based classification techniques and investigate the effect of gender on the writing style. For this purpose, a user document is generated for each user by concatenating all outgoing messages of the user. Each user document forms a classification instance whose features are defined by the information within the user document. Two different techniques

are investigated for classifying the users (i.e., their documents) according to the gender: term-based classification and style-based classification. In the first approach, the set of features is taken as the set of distinct words in the user document. In the second approach, the stylistic properties of a user document are incorporated as its features. For this purpose, various stylistic features, including some well-known features used in literature [15] as well as some newly proposed features, are extracted from the message logs and used with the hope of improving the classification accuracies.

Although the chat dataset used in this work is completely textual, the style of chat messages is quite different than that of any other textual data used in literature. First, the use of punctuation marks varies widely for each user. Some users omit punctuation marks in their messages while some overuse them. Second, since conversations occur in real-time and there is no medium for communication other than text, computer-mediated communication has its own means for transferring emotions. Smileys and emoticons are the commonly known and widely used means of representing feelings within text. Smileys (e.g., ":-)", ":-(") are the sequences of punctuation marks that represent feelings such as happiness, enthusiasm, anger, and depression. Emoticons (e.g., "Awesomeeee!") are consciously done misspellings that put a greater emphasis on an expression. Since the use of these emotion-carriers are closely related to the writing style of an individual, they provide valuable information about their author. However, the existence of emotion-carriers makes standard methods (e.g., stemming and part-of-speech tagging) used for authorship analysis impractical for chat datasets. In our case, the messages contain only ASCII characters since all non-ASCII Turkish characters are replaced with their ASCII counterparts. Hence, the use of natural language processing techniques is even more restricted. Another difference of chat datasets from other textual material is that the messages have limited length. According to Rudman [15], in order to gather sound information on the writing style of an author, the documents should contain at least 1000 words on a specific subject. Finally, in most of the work in literature, the documents in question are selected over a restricted topic. In chat datasets, each message may have a different topic, resulting in user documents with multiple topics. Using a dataset without a restricted content may bias the classification with respect to the topic of the message instead of the authors' gender.

4 Gender Prediction

4.1 Term-Based Classification

We formulate the gender prediction problem as a text classification problem as follows. In our case, each user document is composed of the words typed by a user. The vocabulary of the documents forms the feature set, and the users (i.e., their documents) correspond to the instances to be classified. There are two class values for an instance: male or female.

Given these, the gender prediction problem can be considered as a single-label, binary text classification problem [16]. A supervised learning solution to

this problem is to generate a prediction function, which will map each user document onto one of the male or female classes. In the rest of the paper, we may use the words "term" and "feature" as well as "user document" and "instance", interchangeably.

The above-mentioned prediction function can be learned by any of the existing supervised classification algorithms via training over a representative set of documents whose authors' genders are known. In order to compute this function, we employ four well-known algorithms (k-NN, naive Bayesian, covering rules, and back propagation), which are widely used in machine learning literature. For an excellent survey about the use of machine learning techniques in text classification, the interested reader may refer to [16].

4.2 Style-Based Classification

Although term-based classification is widely used in literature [3], the results of this approach are biased by the topics of the documents. Another method for representing the author is to employ linguistic preferences of an author. Finding writing habits of an author is known as stylometry. The problem in stylometric studies can be summarized as finding similarities between documents using statistics and deriving conclusions from the stylistic fingerprints of an author [3]. A detailed overview of the stylometric studies can be found in [17]. According to Rudman [15], there are more than 1000 stylistic features that may be used to discriminate an author. However, there is no consensus on the set of best features that represents the style of a document.

In this study, several stylistic features are extracted from the chat dataset and examined in order to find the best representation for the messages written by a user. Word lengths, sentence lengths, and function word usage are well-known and widely applied stylistic features [3]. Word lengths and sentence lengths provide statistical information about the author's documents, and function words describe the sentence organization of an author. In our work, a stopword list of pronouns, prepositions, and conjunctions are used as function words. Analysis

Table 1. The stylistic features used in the experiments

Feature	Description	Possible feature values
message length	average message length	low, medium, high
word length	average word length	low, medium, high
stopword usage	frequency of stopwords	low, medium, high
stopwords	a list of 78 stopwords	exists, not exists
smiley usage	frequency of smileys	low, medium, high
smileys	a list of 79 smileys	exists, not exists
punctuation usage	frequency of punctuation marks	low, medium, high
punctuation marks	a list of 37 punctuation marks	exists, not exists
vocabulary richness	number of distinct words	poor, average, rich
character usage	frequency of each character	low, medium, high

of vocabulary richness is also considered as an important stylistic characteristic of an author. The frequency of distinct words within a document is used to represent the vocabulary richness of an author.

In addition to the traditional stylistic features, this study includes several other stylistic features that may describe authors' stylistic fingerprints in written discourse. Since the messages in question are unedited, punctuation usage can be a discriminating factor between different authors. As a computer-mediated text, the chat messages contain emotion-carriers called smileys and emoticons. In this work, an extensive list containing 79 different smileys is used. The overuse of alphabet characters are traced within each message in order to detect the use of emoticons. A summary of the stylistic features used in this study is given in Table 1.

5 Experiments

5.1 Preprocessing

The imbalance in a dataset may form a crucial problem for text classification [18]. The chat dataset used in the experiments is imbalanced due to the following two reasons. First, the number of male and female users is not equal. To alleviate this problem, undersampling [19] is used to balance the number of male and female instances. Each instance is scored with respect to the total number of words he/she uses, and equal number of instances with highest word count are selected as the best representatives of their respective classes. Second, the total number of distinct words used by each user varies. This variance is alleviated by applying a windowing mechanism for each instance. A fixed number of consecutive words are selected from each user document, and the remaining words are discarded.

The high dimensionality of our chat dataset is another factor which badly affects the applicability of machine learning algorithms. Feature selection [20] is a widely used preprocessing step for reducing the dimensionality. In this work, χ^2 (CHI square) statistic is used to calculate the discriminative power of each feature. Experiments are performed using a selected set of the most discriminating features.

5.2 Experimental Setup

A selection of classifiers from the Harbinger machine learning toolkit [21] is used in the experiments. The selected classifiers are k-NN, naive Bayesian, covering rules and back propagation. 10-fold cross-validation is used in all experiments. Each experiment is repeated five times and average accuracy results are reported. The accuracy is defined as the number of instances whose gender is correctly predicted divided by the total number of predictions.

In each experiment, 90% of the most discriminative features are used as representative features. For the text classification tests, a window of 3000 words is selected as the document sample of a user. For the k-NN algorithm, cosine similarity metric is used as the distance metric, and the number of nearest neighbors (k) is set to 10.

Table 2. Accuracies achieved by four different classifiers in predicting the gender of a chat user

	k-NN		Naive Bayesian	
Number of instances	Term-based	Style-based	Term-based	Style-based
25 male–25 female	72.4	56.8	76.0	72.0
50 male–50 female	73.4	63.2	80.0	71.8
100 male–100 female	74.5	60.7	81.5	81.9
200 male–200 female	72.2	62.4	78.2	81.7

	Covering rules		Back propagation	
Number of instances	Term-based	Style-based	Term-based	Style-based
25 male–25 female	49.2	50.4	54.0	71.6
50 male–50 female	53.4	51.2	50.0	75.4
100 male–100 female	58.3	64.2	54.0	80.8
200 male–200 female	56.4	64.9	58.0	79.6

5.3 Results

In this study, it is proposed that the gender of a chat user is distinguishable using the information derived from the messages written by that particular user. In order to test this claim a variety of experiments are done. The experiments are conducted on 100, 200, and 400 users selected from the chat dataset. Table 2 summarizes the accuracy results obtained from the experiments. In term-based classification, naive Bayesian classifier achieves the best results with an accuracy of 81.5%. In general, as the number of instances increases, the use of stylistic features performs better than term-based classification. In style-based classification, naive Bayesian and back propagation perform well with similar accuracies of 81.9% and 80.8%, respectively.

Table 3 shows the effect of feature selection on classification accuracy. The tests are done using the naive Bayesian classifier over a set of 200 users. In order to emphasize the effect of feature selection, a shorter window size of 800 words is used as the document for each user. As the feature space in the style-based classification tests is much smaller than that of the term-based classification

Table 3. Effect of feature selection on classification accuracies

Feature selection	Term-based	Style-based
1%	70.5	60.2
5%	73.4	65.2
25%	76.2	72.6
50%	74.6	77.2
60%	73.8	78.6
70%	75.2	78.8
80%	75.7	78.9
90%	74.7	81.8

Table 4. Effect of discarding a stylistic feature on classification accuracy

Discarded feature	k-NN	Naive Bayesian	Covering rules	Back propagation
none	62.5	81.2	63.7	79.5
message length	60.7	80.7	62.2	80.8
word length	61.6	80.6	63.9	79.4
stopwords	64.6	83.3	67.8	81.9
smileys	59.2	80.2	60.8	80.2
punctuation marks	63.4	75.6	62.4	76.5
vocabulary richness	61.8	81.8	60.7	79.5
character usage	61.8	81.2	62.5	79.6
best feature set	64.6	84.2	68.2	82.4

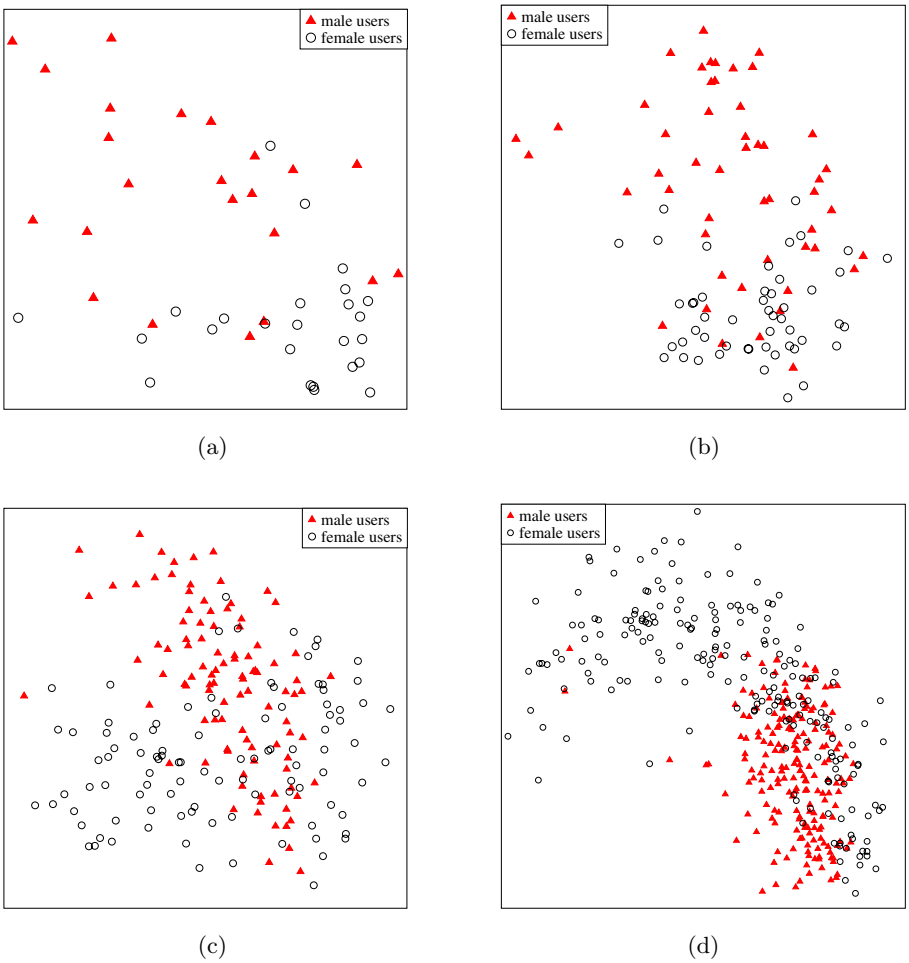

Fig. 1. Distribution of gender according to the principal component analysis for (a) 50, (b) 100, (c) 200, and (d) 400 users, where each case contains an equal number of males and females

Table 5. The list of the most discriminating words and their respective χ^2 values for male and female users

Male		Female	
Word	χ^2	Word	χ^2
abi (brother)	121.3	ayy (boy!)	9.6
olm (buddy)	94.2	kocam (my husband)	7.6
lazim (required)	80.9	okulum (my school)	7.6
tane (amount)	80.3	sevgilimin (my lover's)	7.1
kac (how many)	78.5	suanda (at the moment)	6.6
olmaz (no way)	74.5	iyilik (fine)	6.6
var (exists)	72.0	mersi (thanks)	6.4
baba (fellow)	70.9	byeeee (byeeee)	6.4
IP (IP)	66.5	beeen (I)	6.1
biri (someone)	65.9	kocama (to my husband)	6.1

tests, discarding a percent of least important features from the instances badly affects the style-based classification relatively more. This is mostly because, as the feature space becomes smaller, instances also become similar to each other, and hence classifiers do not function well.

The effectiveness of stylistic features are also questioned in this study. As there is no consensus in literature on the set of the best stylistic features to be used, experiments are conducted in order to evaluate which stylistic features are useful for discriminating the gender of an author. For this purpose, in each experiment, one of the stylistic features is left out, and the accuracy of the classifier is re-evaluated. Table 4 displays the results of the experiments conducted on a selection of 200 users. According to Table 4, the k-NN classifier works best when the stopwords are left out; the back propagation classifier works best when average message lengths, stopwords, the use of smileys, vocabulary richness, and over-used character frequencies are left out; the covering rules classifier achieves its best results using a feature set without stopwords and word lengths. Among the four classifiers, the naive Bayesian classifier achieves the best accuracy (84.2%) using a feature set without stopwords and vocabulary richness measures.

In order to visualize the predictability of gender, principle component analysis (PCA) is used. In this technique, each user document is represented with a vector generated using the distinct words in the user document, and the dimensionality of this vector is reduced using PCA. Figure 1 shows the PCA results for datasets of varying size. It is important to note that the values on the data points are not displayed since they are not indicative of anything. Only the relative proximities of the data points are important. The results clearly show that the use of words in chat messages can be used to discriminate the gender of a user.

6 Concluding Discussion

In this paper, the predictability of the genders of the users involved in computer-mediated conversations is questioned. The word selection and message organization of many chat users are examined. Experiments are conducted in order to

predict the gender over a large, real-life chat dataset. The experimental results led to the finding that both word usage and writing habits of users of different sex vary significantly. Table 5 shows a sample set of discriminative words along with their χ^2 values. It is apparent that males tend to produce more decisive and dominating words that can be considered as slang. On the other hand, female conversations involve more possessive and content-dependent words. Also, the use of emoticons and smileys are distinguishing characteristics of the female writing style.

The stylometric analysis also provides interesting results. In general, female users tend to prefer to use longer and content-bearing words. They also prefer to organize shorter sentences than male users and omit stopwords and punctuation marks. The use of smileys and emotion-carrier words is more common in female users than male users. Long chat messages and the use of short words are the most discriminating stylistic features of male users. Also, the use of stopwords and punctuation marks widely varies for male users. They use punctuation marks and stopwords either heavily or very lightly.

References

1. Love, H.: Attributing Authorship: An Introduction. Cambridge University Press (2002)
2. Corney, M.W.: Analysing E-mail Text Authorship for Forensic Purposes. M.S. Thesis. Queensland University of Technology (2003)
3. Holmes, D.I.: Analysis of Literary Style - A Review. Journal of the Royal Statistical Society **148**(4) (1985) 328–341
4. Elliot, W.E.Y., Valenza, R.J.: Was the Earl of Oxford the True Shakespeare? A Computer Aided Analysis. Notes and Queries **236** (1991) 501–506
5. Merriam, T., Matthews, R.: Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. Literary and Linguistic Computing. **9** (1994) 1–6
6. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist. Addison-Wesley (1964)
7. Holmes, I., Forstyh, R.: The Federalist Revisited: New Directions in Authorship Attribution. Literary and Linguistic Computing **10**(2) (1995) 111–127
8. Tweedie, F.J., Singh, S., Holmes, D.I.: Neural Network Applications in Stylometry: The Federalist Papers. Computers and the Humanities **30**(1) (1996) 1–10
9. Patton, J. M., Can, F.: A Stylometric Analysis of Yasar Kemal's Ince Memed Tetralogy. Computers and the Humanities **38**(4) (2004) 457–467
10. Graham, N., Hirst, G., Marthi, B.: Segmenting Documents by Stylistic Character. Natural Language Engineering **11**(4) (2005) 397–415
11. Vel, O. de, Corney, M., Anderson, A., Mohay, G.: Language and Gender Author Cohort Analysis of E-mail for Computer Forensics. In: Second Digital Forensics Research Workshop. (2002)
12. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. Literary & Linguistic Computing **17**(4) (2002) 401–412
13. Kessler, B., Nunberg, G., Schutze, H.: Automatic Detection of Text Genre. In: Proceedings of the 35th Annual Meeting on Association for Computational Linguistics. (1997) 32–38

14. Spafford, E.H., Weeber, S.A.: Software Forensics: Can We Track Code to Its Authors? Computers and Security **12** (1993) 585–595
15. Rudman J.: The State of Authorship Attribution Studies: Some Problems and Solutions. Computers and the Humanities **31**(4) (1998) 351–365
16. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys **34**(1) (2002) 1–47
17. Holmes, D.I.: Authorship Attribution. Computers and the Humanities **28**(2) (1994) 87–106
18. Liu, A.Y.C.: The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets. M.S. Thesis. University of Texas at Austin (2004)
19. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Data Sets: One-sided Sampling. In: Proceedings of the Fourteenth International Conference on Machine Learning. (1997) 179–186
20. Yang., Y, Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. (1997) 412–420
21. Cambazoglu, B.B., Aykanat, C.: Harbinger Machine Learning Toolkit Manual. Technical Report BU-CE-0503, Bilkent University, Computer Engineering Department, Ankara (2005)