Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 44 (2008) 1448–1466

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Chat mining: Predicting user and message attributes in computer-mediated communication

Tayfun Kucukyilmaz [a], B. Barla Cambazoglu [b], Cevdet Aykanat [a,*], Fazli Can [a]

[a] *Computer Engineering Department, Bilkent University, TR 06800 Bilkent, Ankara, Turkey*
[b] *Yahoo! Research Barcelona, Barcelona, Spain*

Received 1 August 2007; received in revised form 14 December 2007; accepted 29 December 2007
Available online 4 March 2008

Abstract

The focus of this paper is to investigate the possibility of predicting several user and message attributes in text-based, real-time, online messaging services. For this purpose, a large collection of chat messages is examined. The applicability of various supervised classification techniques for extracting information from the chat messages is evaluated. Two competing models are used for defining the chat mining problem. A term-based approach is used to investigate the user and message attributes in the context of vocabulary use while a style-based approach is used to examine the chat messages according to the variations in the authors' writing styles. Among 100 authors, the identity of an author is correctly predicted with 99.7% accuracy. Moreover, the reverse problem is exploited, and the effect of author attributes on computer-mediated communications is discussed.

Keywords: Authorship analysis; Chat mining; Computer-mediated communication; Machine learning; Stylistics; Text classification

1. Introduction

With the ever-increasing use of the Internet, computer-mediated communication via textual messaging has become popular. This type of electronic discourse is observed in point-to-point or multicast, text-based online messaging services such as chat servers, discussion forums, email and messaging services, newsgroups, and IRCs (Internet relay chat). These services constantly generate large amounts of textual data, providing interesting research opportunities for mining such data. We believe that extracting useful information from this kind of messages/conversations can be an important step towards improving the human–computer interaction.

According to a study by Jonsson (1998), "electronic discourse is neither writing nor speech, but rather written speech or spoken writing, or something unique". Due to its mostly informal nature, electronic discourse

* Corresponding author. Tel.: +90 312 290 1625; fax: +90 312 266 4047.
E-mail addresses: ktayfun@cs.bilkent.edu.tr (T. Kucukyilmaz), barla@yahoo-inc.com (B.B. Cambazoglu), aykanat@cs.bilkent.edu.tr (C. Aykanat), canf@cs.bilkent.edu.tr (F. Can).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2007.12.009

Table 1
The summary of abbreviations

AA	Authorship attribution	*k*-NN	k-Nearest neighbor
AC	Authorship characterization	NB	Naive Bayesian
CE	Cross entropy	NN	Neural networks
DA	Discriminant analysis	PCA	Principal component analysis
DT	Decision trees	PRIM	Patient rule induction method
EG	Exponentiated gradient	RM	Regression models
GA	Genetic algorithms	SD	Similarity detection
HMM	Hidden Markov models	SVM	Support vector machines
IRC	Internet relay chat	TC	Text classification

has major syntactic differences from discourse in literary texts (e.g., word frequencies, use of punctuation marks, word orderings, intentional typos). The informal nature of electronic discourse makes the information obtained more realistic and reflects the author attributes more accurately. Analysis of electronic discourse may provide clues about the attributes of the author of a discourse and the attributes of the discourse itself.

Specifically, machine learning can be a powerful tool for analyzing electronic discourse data. This work particularly concentrates on the data obtained from chat servers, which provide a point-to-point, online instant messaging facility over the Internet. We investigate the rate of success in the problem of predicting various author- and message-specific attributes in chat environments using machine learning techniques. For this purpose, we first employ a term-based approach and formulate the chat mining problem as an automated text classification problem, in which the words occurring in chat messages are used to predict the attributes of the authors (e.g., age, gender) or the messages (e.g., the time of a message). Second, we employ a style-based approach and investigate the effect of stylistic features (e.g., word lengths, use of punctuation marks) on prediction accuracies, again for both author and message attributes. Finally, we briefly discuss the effect of the author and message attributes on the writing style.

The main contributions of this study are four-fold. First, the chat dataset used in this work has unique properties: the messages are communicated between two users; they are unedited; and they are written spontaneously. We believe that extracting information from real-time, peer-to-peer, computerized messages may have a crucial impact on the areas such as financial forensics, threat analysis, and detection of terrorist activities in the near future. Our work presents a new effort in that direction, aiming to retrieve previously unexplored information from computerized communications. Second, for the first time in the literature, several interesting attributes of text and its authors are examined. Examples of these attributes are educational affiliations and connectivity domains of the authors and the receivers of the messages. Third, the performance of term- and style-based feature sets in predicting the author and message attributes are compared via extensive experimentation. Fourth, to the best of our knowledge, our work is the first one that investigates real-time, peer-to-peer, computerized communications in the context of authorship studies. Our findings are good pointers for researchers in this new application area, namely chat mining.

The rest of the paper is organized as follows. Table 1 displays a list of frequently used abbreviations in this paper. We provide a detailed literature survey of the related work in Section 2. In Section 3, we discuss the characteristics of computer-mediated communication environments and elaborate on the information that can be extracted from such environments. Section 4 introduces the chat mining problem and discusses our formulations, which are based on the use of term- and style-based feature sets. In Section 5, we provide information about the dataset used in this study and present our framework for solving the chat mining problem. Section 6 provides the results of a large number of experiments conducted to evaluate the feasibility of predicting various author and message attributes in a chat environment. In Section 7, we finalize the paper with a concluding discussion.

2. Related work

In the last 10 years, the Internet has become the most popular communication medium. Chat servers, IRCs, and instant messaging services provide online users the ability to communicate with each other simulta-

neously. Discussion forums, emails, and newsgroups enable their users to create virtual communities regardless of geographical and political barriers. This information dissemination platform provides new research possibilities such as assessing the task-related dimensions of the Internet use. In their work, Dickey, Burnett, Chudoba, and Kazmer (2007) examine the communication process of chat users in an industrial setting. They investigate how customers and customer service representatives respond to each other and identify the reasons of miscommunication between partners. The collaborative work within virtual groups is explored by Walther, Bunz, and Bazarova (2005). The authors identify six communication rules for enhancing trust, which in turn enable chat users to work more efficiently. Radford (2005) examines several problems concerning communications in a virtual library reference service. The quality of chat encounters between librarians and clients, compensation of lack of emotional cues, and relational dimensions of chat references are among the questions investigated. The author identifies several relational facilitators as well as communication themes and concludes that computer-mediated communication is no less personal than face-to-face communication.

Understanding the user behavior is another aspect of the ongoing research on computer-mediated communication. Radford and Connaway (2007) examine the communication and information seeking preferences of the Internet users. They also compare traditional libraries and the Internet as the means for an information repository and emphasize the fact that the Internet is starting to become an alternative for text-based communication.

The investigation of chat user attributes is another dimension that attracts researchers. Herring and Paolillo (2006) examine gender variations in Web logs using logistic regression techniques. However, the authors cannot find any conclusive results binding the users' genders and Web writings. In their work, Herring and Danet (2007) examine several aspects of the language use in the Internet. They assert that gender is reflected in online discourse in every language they studied.

Extracting interesting information from anonymous electronic document collections using authorship attribution may also provide several research opportunities. A quick literature survey reveals the fact that the previous studies in authorship attribution were mostly considered in the context of law enforcement (Turell, 2004), religious studies (Pollatschek & Radday, 1981; Sabordo, Chai, Berryman, & Abbott, 2005), and humanities (Can & Patton, 2004; Elliot & Valenza, 1991; Mosteller & Wallace, 1964). In the past few years, the examination of electronic discourse in the context of authorship studies started to get attention of a growing number of researchers.

The history of authorship studies dates back to more than two millennia. The first work in literature is reported in the fourth century BC, when the librarians in the famous library of Alexandria studied the authentication of texts attributed to Homer (Love, 2002). Since then, a large number of documents have been the focus of authorship studies. Broadly, the authorship studies in literature can be divided into three categories (Corney, 2003; Zheng, Li, Chen, & Huang, 2006): authorship attribution, similarity detection, and authorship characterization.

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 (Elliot & Valenza, 1991; Hota, Argamon, & Chung, 2006; Merriam & Matthews, 1994) and the identification of the authors of the disputed Federalist Papers (Holmes & Forsyth, 1995; Levitan & Argamon, 2006; Mosteller & Wallace, 1964; Tweedie, Singh, & Holmes, 1996). Similarity detection aims to find the variations in the writing style of an author (Patton & Can, 2004) or to find the resemblances between the writings of different authors, mostly for the purpose of detecting plagiarism (Graham, Hirst, & Marthi, 2005).

Authorship characterization is the task of assigning the writings of an author into a set of categories according to the author's sociolinguistic attributes. Some attributes previously investigated in literature are gender (Koppel, Argamon, & Shimoni, 2002; Kucukyilmaz, Cambazoglu, Aykanat, & Can, 2006; Vel, Corney, Anderson, & Mohay, 2002), language background (Vel et al., 2002), and education level (Juola & Baayen, 2005). Koppel et al. (2002), and Kucukyilmaz et al. (2006) evaluated methods for determining the gender of a document's author. Vel et al. (2002), in addition to gender, tried to predict the language background of authors using machine learning techniques. Juola and Baayen (2005) analyzed the educational backgrounds of the authors employing cross entropy.

With the advent of computers, it has become possible to employ sophisticated techniques in authorship analysis. The techniques employed in authorship analysis can be broadly categorized as statistical and

machine learning techniques. Examples of statistical techniques are Hidden Markov models (Khmelev & Tweedie, 2001), regression models (Kessler, Nunberg, & Schutze, 1997), cross entropy (Juola & Baayen, 2005), discriminant analysis (Can & Patton, 2004; Karlgren & Cutting, 1994; Krusl & Spafford, 1997; Thomson & Murachver, 2001), and principle component analysis (Baayen, van Halteren, & Tweedie, 1996; Burrows, 1987; Holmes, 1994). Machine learning techniques are also frequently used in authorship studies. Most commonly used techniques are *k*-nearest neighbor (Krusl & Spafford, 1997; Kucukyilmaz et al., 2006; Stamatos, Fakotakis, & Kokkotakis, 2000), naive Bayesian (Kjell, 1994; Kucukyilmaz et al., 2006; Stamatos et al., 2000), support vector machines (Corney, 2003; Corney, Anderson, Mohay, & Vel, 2001; Joachims, 1998; Tsuboi & Matsumoto, 2002; Zheng et al., 2006), genetic algorithms (Holmes & Forsyth, 1995), decision trees (Zheng et al., 2006), and neural networks (Graham et al., 2005; Kjell, 1994; Krusl & Spafford, 1997; Kucukyilmaz et al., 2006; Merriam & Matthews, 1994; Stamatos et al., 2000; Tweedie et al., 1996; Zheng et al., 2006).

With the widespread use of computers, new pursuits that reflect the personal characteristics of individuals drew attention of authorship studies. Computer programming and musical composition are examples of such pursuits. Spafford and Weeber (1993) and Krusl and Spafford (1997) used several structural and syntactic features to predict the author of a program. They generate these features by analyzing the variations in programming construct preferences of the authors. The work of Spafford and Weeber (1993) achieved 73% accuracy in predicting the author of 88 programs written by 29 different authors. In their work, Backer and Kranenburg (2004) analyzed the musical style of five well-known composers using various classification algorithms on a dataset with computer-generated features like the stability measures of the composition, voice density, and entropy measures.

The emergence of electronic discourse also presents interesting opportunities for authorship analysis. As electronic discourse becomes a popular form of communication, detecting illegal activities by mining electronic discourse turns out to be important. In their work, Vel et al. (2002) analyzed the information in email messages in order to identify the distinguishing features in writing styles of emails for predicting authors' identity, gender, and language background. In addition to some well-known stylistic features, they used features like smileys and emoticons. They achieved 72.1% and 85.6% accuracies in predicting the gender and language background of more than 300 authors, respectively.

Table 2
A summary of the previous works on authorship analysis

Study	Type	Technique	Features
Mosteller and Wallace (1964)	AA	Statistics	Style
Burrows (1987)	AA	PCA	Style
Elliot and Valenza (1991)	AA	Statistics	Both
Karlgren and Cutting (1994)	TC	DA	Style
Kjell (1994)	AA	NB, NN	Style
Merriam and Matthews (1994)	AA	NN	Style
Holmes and Forsyth (1995)	AA	GA	Style
Baayen et al. (1996)	AA	PCA	Both
Kessler et al. (1997)	TC	RM	Style
Krusl and Spafford (1997)	AA	*k*-NN, DA, NN	Style
Joachims (1998)	TC	SVM	Term
Stamatos et al. (2000)	AA, TC	*k*-NN, NB, NN	Style
Khmelev and Tweedie (2001)	AA	HMM	Term
Thomson and Murachver (2001)	AC	DA	Style
Tsuboi and Matsumoto (2002)	AA, AC	SVM	Style
Vel et al. (2002)	AC	SVM	Term
Argamon et al. (2003)	AA	EG	Style
Corney (2003)	AA	SVM	Style
Patton and Can (2004)	AC	DA	Style
Graham et al. (2005)	SD	NN	Style
Juola and Baayen (2005)	AA, AC	CE	Term
Hota et al. (2006)	AC	SVM	Style
Kucukyilmaz et al. (2006)	AC	*k*-NN, NB, NN	Both
Zheng et al. (2006)	AA	SVM, DT, NN	Style

Tsuboi and Matsumoto (2002) analyzed email messages for predicting the identity of their authors using a term-based feature set. Thomson and Murachver (2001) analyzed the gender of a number of email authors and concluded that email authors had used gender-preferential language in informal electronic discourse. Zheng et al. (2006) constructed a language-independent framework to predict the identity of the author of online Chinese and English newsgroup messages. For a selection of 20 authors they have succeeded in predicting the identity of the authors with an impressive 95% accuracy for the English message collection and 88% accuracy for the Chinese message collection. Argamon, Saric, and Stein (2003) also studied newsgroup messages for identification of the authors using a style-based classification approach. Although they used a highly imbalanced dataset, over 40% accuracy is achieved in predicting the messages of 20 different authors.

Zheng et al. (2006) presented a table that provides a summary (features used, type of analysis, and dataset properties) of the previous works on authorship analysis. Here, we provide a similar table with additional information for a number of previous works. In chronological order, Table 2 gives details such as the analysis techniques used in the works and the type of the features used (i.e., term-based or style-based features). In compliance with our previous taxonomy, the table categorizes each work as an authorship attribution (AA), similarity detection (SD), or authorship characterization (AC) task. Several text classification (TC) works, which are closely related with authorship studies, are also displayed in the table.

3. Computer-mediated communication

3.1. Characteristics

Using textual messages in order to interact with other people is a popular method in computer-mediated communication. Point-to-point instant messaging, also referred to here as chatting, has several properties which makes it unique with respect to both literary writing and messaging in other types of online services: messages (1) are written by users with a virtual identity; (2) specifically target a single individual; (3) are unedited; and (4) have a unique style and vocabulary. Below, we elaborate more on these characteristics.

In most chat servers, the real identity of a user is hidden from other users by a virtual identity, called "nickname". Typically, the users have the option of building up this virtual identity and setting its different characteristic features. This gives the users the opportunity to provide others false information about their real identities. For example, a male user may set the gender of his virtual identity as female and try to adapt his writing style accordingly to fool others. Having such misleading information in chat environments makes authorship attribution and characterization quite difficult even for domain experts.

Unlike literary writing, where the documents are written for public audience, chat messages target a particular individual. Most often, chat messages are transmitted between two users, that is, each message has a specific sender and a receiver. The writing style of a user not only varies with his personal traits, but also heavily depends on the identity of the receiver. For example, a student may send a message to another student in a style which is quite different from the style of a message he/she writes to his supervisor. This type of an ability of effectively changing one's writing style is known as sociolinguistic awareness (Hakuta, 1991). As an interesting genre detection task, chat messages can be examined in order to find out who the receiver is.

Books and plays are the most common type of literary material used in authorship analysis (Foster, 2000). This type of documents are usually modified by editors who polish the initial drafts written by authors. Hence, most of the time, the writing style of the original author is mixed with that of an editor. Rudman (1998) discusses the undesirable effects of this type of editing on authorship analysis and concludes that edited texts are hard to mine since stylistic traces of the author and the editor are not separable. The real-time nature of chat messages prevents any editorial changes in electronic discourse, and thus the writing style reflects that of the original author. In this aspect, it is quite valuable to work on unedited chat messages. However, in the mean time, having no editorial modifications means that, in chat messages, misspellings are more frequent compared to edited text. It is debatable whether these misspellings are part of an author's writing style or not.

Due to its simultaneous nature, electronic discourse reflects the author's current emotional state much better than any other writing. Since the messages transmitted between users are purely textual, chat messaging has evolved its own means for transferring emotions. Emoticons (emotion icons) are commonly known and widely used ways of representing feelings within computer-mediated text (Wikipedia, 2007). We restrict our

work on a particular subset of emoticons: smileys. Smileys, (e.g, ":-)" and ":-(") are sequences of punctuation marks that represent feelings such as happiness, enthusiasm, anger, and depression. Repetition of specific characters in a word can also be used as a means of transferring emotions by putting an emphasis on a text. (e.g. "Awesomeeee!"). In chat messages, the use of such consciously done misspellings is also frequent. Since the use of smileys and emphasized words is highly dependent on the writing style of an author, they pose valuable information. However, preserving such information makes traditional text processing methods (e.g., stemming and part of speech tagging) unsuitable for mining chat messages.

3.2. Predictable attributes

In general, chat messages can be used to predict two different types of attributes: user- or message-specific attributes. In the first type, the distinguishing features of a chat message may be used to predict the biological, social, and psychological attributes of the author who wrote the message. In the latter, the distinguishing features may be used to predict the attributes of the message itself.

Examples of user-specific attributes are gender, age, educational background, income, linguistic background, nationality, profession, psychological status, and race. In this work, we concentrate on four different user-specific attributes: gender, age, educational environment, and Internet connection domain of the users. Among these attributes, the gender of an author is widely examined in literature (Kucukyilmaz et al., 2006; Vel et al., 2002), and it is observed that authors have the habit of selecting gender-preferential words (Thomson & Murachver, 2001). In this work, we also try to predict the user age based on the fact that every generation has its own unique vocabulary. Predicting the age of a user may be useful for profiling the user and hence may help in forensic investigations. Educational environment is also worth studying since it is possible that the vocabulary and writing style of a user might be affected by the school he/she is affiliated with. In order to test this claim, we analyzed the chat messages of users in different universities. We also noted that computer-mediated communication adds new dimensions whose analysis may yield valuable information. As an illustrative task, we try to predict the Internet connection domains of users, which may have veiled means for the educational and occupational status of a user. For example, a user connected from the ".edu" domain probably has an affiliation with a university, whereas a user connected from the ".gov" domain possibly works for the government.

For message-specific attributes, we concentrate on three attributes: author, receiver, and time of the messages. The identity of the author of a given text is the most frequently studied attribute in authorship analysis (Holmes & Forsyth, 1995; Krusl & Spafford, 1997; Mosteller & Wallace, 1964; Zheng et al., 2006). In case of chat mining, the characteristics of chat messages are firmly attached to the author's linguistic preferences. Hence, we try to predict the authors of chat messages as a typical authorship attribution task. The audience of a chat message may also affect the lingual preferences of an author. For the first time in literature, we try to predict the audience of textual documents; i.e., the receivers of the chat messages. The real time nature of chat messages makes it possible to examine whether the time a message is written is predictable. For example, in active hours of the day (morning and afternoon), people may compose long and complex sentences although, in passive hours (evening and night), people may tend to create short and simple sentences. Hence, in this work, we also investigate the predictability of the periods of the day in which chat messages are written.

Table 3 presents a complete list of the attributes we try to predict in this paper. In this table, the number of classes refers to the maximum number of possible values an attribute can have. For example, the gender attri-

Table 3
The attributes predicted in this work and the number of classes available for each attribute

User-specific attributes	Number of classes	Message-specific attributes	Number of classes
Gender	2	Receiver	1165
Age	17	Author identity	1616
School	60	Day period	4
Connection domain	7	–	–

bute has two possible class values (male and female) while the connection domain attribute has seven possible class values, each of which represents a different Internet connectivity domain.

4. Chat mining problem

The chat mining problem can be considered as a single-label classification problem. If the attribute to be predicted is user-specific, a supervised learning solution to this problem is to generate a prediction function, which will map each user instance onto one of the attribute classes. The prediction function can be learned by training supervised classification algorithms over a representative set of user instances whose attributes are known. In case of message-specific attributes, the process is similar. However, this time, the individual chat messages are the instances whose attributes are to be predicted, and the training is performed over a set of chat messages whose attributes are known.

In predicting the user-specific attributes, each user instance is represented by a set of features extracted from the messages that are generated by that particular user. Similarly, in predicting message-specific attributes, each message instance is represented by a set of features extracted from the message itself. In this work, for predicting both types of attributes, we evaluate two competing types of feature sets: term-based features versus style-based features.

When term-based features are used, the vocabulary of the message collection forms the feature set, i.e., each term corresponds to a feature. In predicting user-specific attributes, the set of terms typed by a user represents a user instance to be classified. In predicting message-specific attributes, the terms in a message represent a message instance. This type of a formulation reduces the chat mining problem to a standard text classification problem (Sebastiani, 2002).

In literature, term-based feature sets are widely used (Lam, Ruiz, & Srinivasan, 1999). Unfortunately, term-based features may not always reflect the characteristics of an author since the terms in a document are heavily dependent on the topic of the document. In chat mining, a feature set that is independent of the message topic may lead to better results in predicting the user- and message-specific attributes. Hence, using the stylistic preferences instead of the vocabulary emerges as a viable alternative.

Rudman (1998) states that there are more than 1000 different stylistic features that can be used to define the literary style of an author. The most commonly used stylistic features are word frequencies; sentence and word lengths; and the use of syllables, punctuation marks, and function words (Holmes, 1985). So far, there is no consensus on the set of the most representative features.

This study, in addition to the traditional stylistic features, considers several new and problem-specific stylistic features (e.g., smileys) in order to find better representations for user or message instances. The smileys are important features that are frequently found in chat messages. A summary of the style-based features used in this study is given in Table 4. The stylistic features used in this work are grouped into 10 categories. Each category contains one or more features with categorical feature values. For example, the average word length feature can possibly have three values: short, medium, and long. This discretization is performed depending on the feature value distributions over a set of messages randomly selected from the chat dataset.

Table 4
The stylistic features used in the experiments

Feature category	Features in the category	Possible feature values
Character usage	Frequency of each character	Low, medium, high
Message length	Average message length	Short, average, long
Word length	Average word length	Short, average, long
Punctuation usage	Frequency of punctuation marks	Low, medium, high
Punctuation marks	A list of 37 punctuation marks	Exists, not exists
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
Vocabulary richness	Number of distinct words	Poor, average, rich

5. Dataset and classification framework

5.1. Dataset

The chat dataset used in this paper is obtained from a currently inactive chat server called Heaven BBS, where users had peer-to-peer communication via textual messages. The outgoing chat messages (typed in Turkish) of 1616 unique users is logged for a one-month period in order to generate the dataset. The messages are logged without the notice of the users, but respecting the anonymity of messages. The vocabulary of the dataset contains 165,137 distinct words. There are 218,742 chat messages, which are usually very short (6.2 words per message on the average). The message log of a typical user contains around 160 chat messages.

The dataset also contains users' subscription information such as the name, gender, email address, and occupation. Some fields of the subscription information may be missing as they are optionally supplied by the users. Also, against our best efforts to validate the correctness of the entries, there may be fakes or duplicates among the users.

5.2. Classification framework

In this section, we provide an overview of the framework we developed for solving the chat mining problem. Here, we restrict our framework to prediction of user-specific attributes using the term-based feature set. Extensions of this framework to the message-specific attributes and the style-based feature set are discussed later in this section. Fig. 1 summarizes the classification procedure used in predicting the user-specific attributes. The framework consists of three stages: data acquisition, preprocessing, and classification. The last two stages contain several software modules that execute in a pipelined fashion.

The corpus creation module of the data acquisition stage forms a tagged corpus from the raw message logs obtained from the chat server. In Fig. 2, we provide a sample fragment from this corpus. For each user instance in the corpus, between an "INSTANCE" tag pair, the attributes of the user and the messages typed by the user are stored. The target users receiving the messages of the user are separated by the "RECEIVER" tag pairs. Each receiver may receive multiple messages, which are separated by the "X" tag pairs.

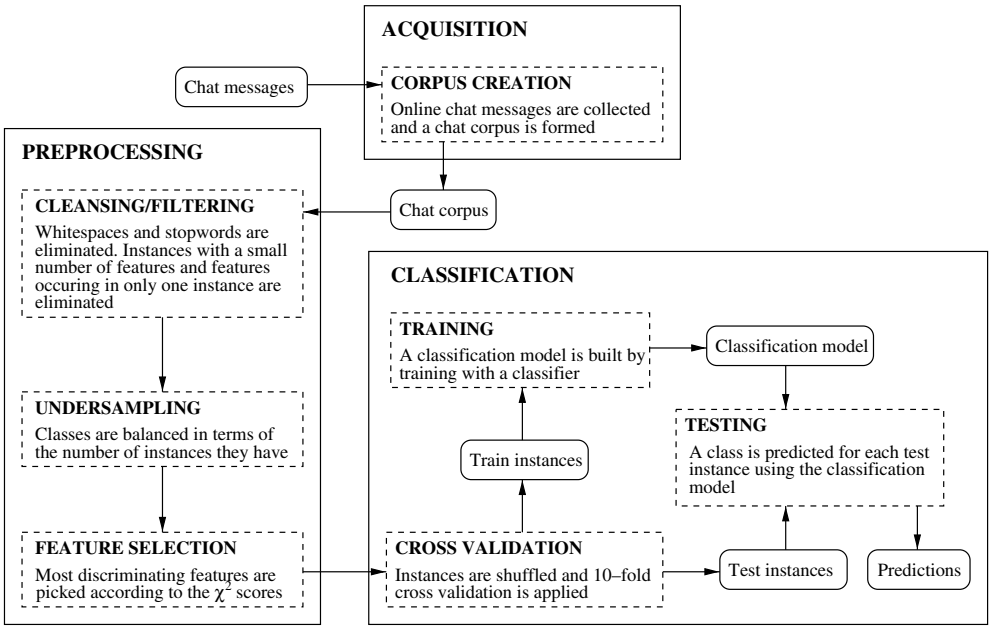

Fig. 1. The classification framework.

```
<INSTANCE=ali>
<NAME=ali guney>
<GENDER=male>
<EMAIL=Guney@alpha.eng.ege.edu.tr>
<DOMAIN=edu>
<SCHOOL=ege>
<BIRTHDAY=19>
<BIRTHMONTH=October>
<BIRTHYEAR=1979>
<HOROSCOPE=libra>
<RECEIVER=blandinka>
<X>
<DATE=Wed Apr 5 16:09:40 2000>
MESELA COK GENIS BIR INSANSIN AMA BAZEN COK KUCUK BIR SEYE
TAKIYOSUN GIBI
//For Example, you are a flexible person.  But Sometimes you concentrate
on small things
</X>
</RECEIVER>
<RECEIVER=ageofeye>
<X>
<DATE=Wed Apr 5 16:10:48 2000>
KONUSMAK ISTIYORMUSUN BENLE
// do you want to talk with me
</X>
</RECEIVER>
</INSTANCE>
```

Fig. 2. A sample fragment of the chat corpus formed. The user name is modified to preserve the anonymity. English translations are added for convenience.

After the chat corpus is generated, it undergoes several preprocessing steps to improve classification accuracies. Each preprocessing step is designed as a separate software module. In our framework, the preprocessing stage involves three modules: cleansing/filtering, undersampling, and feature selection.

The cleansing/filtering module aims to obtain a set of representative terms for each user. For this purpose, non-alphanumeric characters (e.g., whitespaces, punctuation marks) are eliminated. A list of 78 Turkish stopwords (i.e., connectives, conjunctions, and prepositions) is further used to eliminate content-independent terms. Single-word messages are also ignored since these are mostly uninformative salutations. The features of the user instances are formed by the remaining terms, where the tf–idf (term frequency–inverse document frequency) values (Salton & McGill, 1983) are used as the feature values. Finally, the user instances that contain only a small number of features, i.e, those that have less than a pre-determined number of terms, are eliminated.

The existence of imbalanced classes is a crucial problem in text classification (Kubat & Matwin, 1997). If the number of instances selected from each class are not roughly equal, the classifiers may be biased, favoring more populated classes. The main goal of the undersampling module is to balance the number of instances in each class. For this purpose, an equal number of instances with the highest term counts are selected from each class and the remaining instances are discarded. In this dataset, an imbalance is also observed on instance sizes since the number of distinct terms of each user greatly varies. In order to balance instance sizes, a fixed number of consecutive terms is selected for each user, and the remaining terms are discarded.

The high dimensionality of text datasets badly affects the applicability of classification algorithms. Feature selection (Yang & Pedersen, 1997) is a widely used preprocessing stage for reducing the dimensionality of the datasets. In the feature selection module, we employ the χ^2 (CHI square) statistic for every term in order to calculate their discriminative power. Most discriminative features are selected according to the χ^2 scores and used as the feature set. The remaining less discriminative features are eliminated in the feature selection module.

The operation of the modules of the preprocessing stage shows variations in case of message-specific attributes or the style-based feature set. For the case of message-specific attributes, the cleansing/filtering module also employs word blocking. This is because chat messages typically contain only a few words, and it is dif-

ficult to correctly classify a message with this little information. The cleansing/filtering module concatenates multiple consecutive messages of the same user into a single long message. After blocking, the message instances become lengthy enough to have sensible information (Corney, 2003).

In the case of the style-based feature set, instead of terms, a number of stylistic features are extracted. Some of these features contain statistics about the punctuation and stopword usage. Thus, for the construction of style-based feature sets, punctuation marks and stopwords are not eliminated in the cleansing/filtering module. Additionally, for user-specific attributes, the feature sets of all chat messages belonging to a user are combined and used as the feature set for that user. Since the instances contain roughly equal number of features in style-based feature sets, the undersampling module does not try to balance the instance sizes.

The classification stage contains three modules. In the cross validation module, the instances in the dataset are shuffled and divided into 10 equal-sized instance blocks. One of these blocks is selected as the test instance block while other instance blocks are used for the training the framework. The training module uses the training instances supplied by the cross validation module. The output of the training module is a classification model, which is used by the testing module in order to predict the classes of each test instance. The testing module produces a set of predictions based on the classification model and the accuracy of a test is defined as the number of correct predictions divided by the number of total predictions. This operation is repeated 10 times, each time with a different block selected as the test instance block. The average of all predictions gives the prediction accuracy of a classifier. The testing module uses a set of algorithms selected from the Harbinger machine learning toolkit (Cambazoglu & Aykanat, 2005) and SVM-light (Joachims, 1998). An overview of the selected algorithms can be found in the corresponding references.[1]

6. Experimental results

6.1. Experimental setup

In order to examine the predictability of user and message attributes, the personal information within the chat server logs are used. Some attributes, such as the birth year and educational environment, are submitted voluntarily by the users, and hence they may be missing. As a consequence, some attribute classes are very lightly populated and the use of such classes in evaluating the predictability of that attribute may be impractical. Thus, the experiments are conducted on a selection of the most populated classes of each attribute. As an illustrative example, the connectivity domain attribute has seven possible class values. For examining the predictability of the connectivity domain attribute, the most populated two and three classes are selected from the possible seven classes, and the experiments are conducted only on the instances belonging to those classes.

Table 5 summarizes the experiments conducted for estimating the prediction accuracies of each attribute. The table contains information about the number of classes, the number of instances, and a set of sample classes used in each test set. Test sets are tagged by concatenating the attribute name, the number of classes, and the number of instances used to represent each class. For example, the School-3-80 tag corresponds to the experiment conducted for predicting the educational environment of users. This experiment involves three possible classes, each of which contains 80 representative instances. As an example for the case of message-specific attributes, the experiment tagged with Author-10-26 involves 10 possible classes, each of which contains 26 instances. Here, each class represent a different author, and instances correspond to message blocks generated by concatenating a particular author's messages.

A selection of classifiers from the Harbinger machine learning toolkit (Cambazoglu & Aykanat, 2005) is used for predicting the user and message attributes. The selected classifiers are k-NN (Han, Karypis, & Kumar, 2001), NB (McCallum & Nigam, 1998), and PRIM. Additionally, SVM-light (Joachims, 1998) software is used in order to apply SVM to the chat mining problem. In each test setting, 90% of the most discriminative features are used as the representatives. A sequence of 3000 words is used as the maximum document size for term-based feature sets. The remaining terms in the documents containing more than 3000 terms are

[1] The source codes of these algorithms are publicly available online and may be obtained from the following Web addresses: http://cs.bilkent.edu.tr/~aykanat/SoftwarePackages/HMLT.tar.gz; http://download.joachims.org/svm_struct/current/svm_struct.tar.gz.

Table 5
Test sets, their parameters, and sample classes

Test set	Number of instances	Number of classes	Sample classes
Author-2-35	2	35	Andromeda and Taru
Author-10-26	10	26	Andromeda, Taru, Zizer, ...
Author-100-10	100	10	Andromeda, Taru, Zizer, ...
BirthYear-2-30	2	30	Birth year before 1976 (inclusive), Birth year after 1976 (exclusive),
BirthYear-4-30	4	30	1975, 1976, 1977, 1978
DayPeriod-2-34	2	34	Day, night (representing 12-h periods)
DayPeriod-4-17	4	17	Morning, afternoon, evening, night (representing 6-h periods)
Domain-2-35	2	35	.edu, .com
Domain-2-50	2	50	.edu, .com
Domain-2-65	2	65	.edu, .com
Domain-3-30	3	30	.edu, .com, .net
Gender-2-50	2	50	Male, female
Gender-2-100	2	100	Male, female
Gender-2-200	2	200	Male, female
Receiver-2-35	2	35	Andromeda, Taru
Receiver-10-26	10	26	Andromeda, Taru, Zizer, ...
School-2-190	2	190	Bilkent, METU
School-3-80	3	80	Bilkent, METU, Ege
School-3-120	3	120	Bilkent, METU, Ege
School-5-50	5	50	Bilkent, METU, Ege, KHO, ...
School-10-29	10	29	Bilkent, METU, Ege, KHO, ...

discarded. For the k-NN classifier, the cosine similarity measure is used as the distance metric and the number of the nearest neighbors, k, is selected as 10. A polynomial kernel (Joachims, 1998) is used in SVM. Each 10-fold cross-validation experiment is repeated 5 times and the average prediction accuracies are reported.

6.2. Analysis of predictability

In order to visualize the predictability of different attributes, PCA is used. By using PCA, it is possible to reduce the dimensionality of the instances, allowing them to be plotted in two dimensions (Binongo & Smith, 1999). Fig. 3 shows PCA results for four different attributes using a term-based feature set. These attributes are the gender, identity, and Internet connectivity domain of an author and the time period of the messages. As the PCAs of the style-based feature set is similar, they are omitted from this study. Also, note that the coordinate values of the principle component analysis are not displayed. In this work, PCA is only used for the reduction of dimensionality of the dataset. Thus, the values of the data points are not indicative of anything, and only the relative proximities of the data points are important.

Since the data points for each author cover separate regions, it is reasonable to expect high accuracies in predicting the identity of the author of a message. For the PCA of the Internet connection domain, it can be seen that the distribution of data points that belong to the ".com" and ".net" domains cover nearly identical regions while the data points belonging to the ".edu" domain cover a separate region. Hence, it would be reasonable to expect that the ".edu" class could be predicted accurately while ".com" and ".net" domains would be frequently mispredicted. The results of PCA show that it would not be possible to discriminate all attributes equally using a term-based feature set.

6.3. User-specific attributes

Table 6 summarizes the prediction accuracies of the experiments conducted on the user-specific attributes. Among all experiments, the highest prediction rates are achieved for the Internet connection domain of a user.

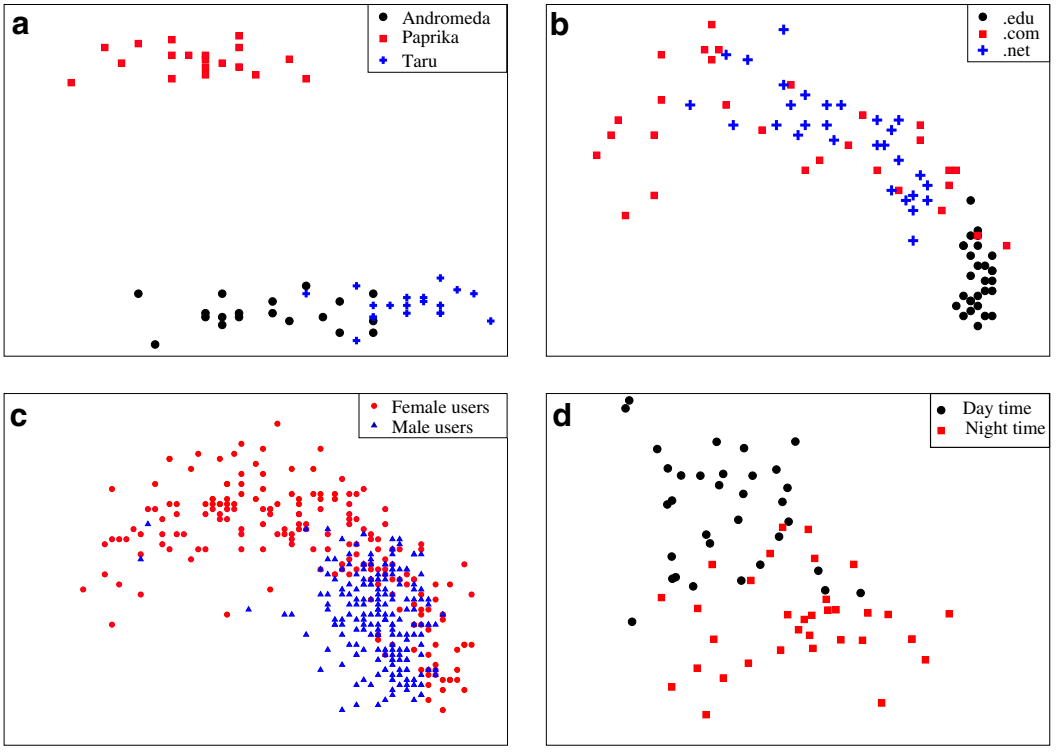

Fig. 3. The results of the PCA for four different attributes (following our earlier convention): (a) Author-3-20; (b) Domain-3-20; (c) Gender-2-200; (d) DayPeriod-2-34.

For this attribute, the NB classifier predicts 91.8% and 68.7% of the test instances correctly for the Domain-2-50 and Domain-3-30 test cases, respectively. The gender, education environment, and the birth year attributes of a user are also predicted accurately. The prediction accuracies of 82.2% and 75.4% are achieved in prediction of the gender and the birth year of a user respectively. The educational environment of a user attains 68.8%, 53.4%, and 39.0% correct prediction rates for the School-2-190, School-5-50, and School-10-29 test

Table 6
Prediction accuracies of experiments conducted on user-specific attributes

	Term-based feature set				Style-based feature set			
Tag	*k*-NN	NB	PRIM	SVM	*k*-NN	NB	PRIM	SVM
BirthYear-2-30	50.1	**60.8**	53.8	56.3	50.0	**75.4**	55.5	48.0
BirthYear-4-30	24.0	**27.3**	20.0	26.5	22.8	**37.4**	19.9	22.0
Domain-2-35	59.7	**90.0**	77.2	64.3	63.9	**90.0**	66.9	59.7
Domain-2-50	58.2	**91.8**	74.0	63.6	64.2	**88.2**	74.4	69.0
Domain-2-65	55.9	**91.4**	79.3	65.2	68.6	**89.8**	78.0	74.1
Domain-3-30	34.0	**67.4**	49.6	39.6	34.7	**68.7**	48.2	45.8
Gender-2-50	73.4	80.0	53.4	**81.5**	63.2	**71.8**	51.2	71.4
Gender-2-100	74.5	81.5	58.3	**82.2**	61.7	**81.9**	64.2	72.3
Gender-2-200	72.2	78.2	56.4	**80.2**	62.4	**81.7**	64.9	77.8
School-2-190	56.8	**68.8**	55.8	66.8	59.3	55.2	50.3	**62.9**
School-3-80	43.6	56.7	35.9	**59.7**	43.1	**47.0**	34.0	51.0
School-3-120	42.7	53.2	41.1	**61.0**	44.1	40.4	32.0	**63.7**
School-5-50	30.8	48.9	26.8	**53.4**	29.1	41.2	25.9	**43.7**
School-10-29	22.5	37.8	17.6	**39.0**	20.4	**26.7**	13.9	26.2

cases respectively. The results of the classification experiments using a term-based feature set lead to the conclusion that gender, identity, and Internet connection domain attributes contain information that reflect the language preferences of a user and it is possible to predict these attributes.

In order to verify whether the experiments are more than some lucky guessing, the level of significance for each experiment is determined. For this purpose, two prediction functions are generated. These functions are used to represent a control group and a treatment group. The control group consists of random guesses for each instance while the treatment group consists of predictions after the classifiers are used. The value of the prediction function is 1 if the instance is correctly predicted and 0 otherwise. Wilcoxon signed-rank test (Wilcoxson, 1945) is used for determining the levels of significance. The significance levels are computed for the best classification result, represented in bold case in Table 6. Table 7 summarizes the z-scores and p-values for each experiment group for user-specific attributes. Noting that the most common level of significance is 5%, all experiments performed significantly better than random guesses. The experiments conducted on the Internet connectivity domain, gender, and educational environment attributes all result in very low levels of significance, which means that the methods proposed in this work can be used effectively to predict these attributes in chat messages.

In predicting the user-specific attributes, the use of term- and style-based feature sets perform almost equally well. While the term-based feature sets performs better than style-based feature sets for predicting the Internet connection domain and the educational environment of a user, the use of style-based feature sets perform better for predicting the birth year of a user.

The performance of different classifiers vary throughout the experiments. The experimental results on the prediction of user-specific attributes show that NB and SVM perform best in all settings although the results show that no single classifier can be the "best performer". While NB performs better than SVM in predicting the connection domain of a user, SVM performs slightly better in predicting the educational environment of a user. k-NN produces the worst results for the prediction of the Internet connection domain while PRIM performs the poorest in prediction of all other attributes. PRIM's poor performance is a result of it being a rule-based classifier. PRIM generates a set of classification rules covering all the instances in a class, and use these rules to classify the test instances. Due to the high dimensionality of the dataset, these rules contain only the most discriminative features, and thus, tend to be valid for only a small subset of the instances in a class. Since such rules fail to classify a large enough subset of the test instances, the classification of PRIM degenerates into random guesses.

Table 7
Significance analysis conducted on user-specific attributes

Tag	Term-based feature set		Style-based feature set	
	z-Score	p-Value	z-Score	p-Value
BirthYear-2-30	1.73	8.3e−1	2.10	2.7e−2
BirthYear-4-30	1.66	1.9e−1	2.03	5.8e−2
Domain-2-35	4.45	6.2e−7	3.74	8.0e−4
Domain-2-50	5.32	9.5e−9	4.27	3.3e−5
Domain-2-65	5.44	8.4e−8	5.61	7.1e−9
Domain-3-30	4.11	3.6e−6	3.94	6.6e−5
Gender-2-50	4.02	3.3e−5	2.32	3.7e−2
Gender-2-100	5.31	3.4e−7	5.39	5.1e−8
Gender-2-200	6.51	6.4e−10	7.11	1.1e−11
School-2-190	3.74	2.0e−4	2.92	3.2e−3
School-3-80	4.72	9.8e−7	2.60	1.2e−3
School-3-120	6.78	1.3e−12	6.64	5.2e−12
School-5-50	7.17	1.1e−12	4.49	2.1e−5
School-10-29	7.10	4.1e−10	3.44	2.1e−5

6.4. Message-specific attributes

Table 8 summarizes the prediction accuracies of experiments conducted on the message-specific attributes. The identity of the author is predicted with perfect accuracy for two and 10 authors using term-based feature sets. The prediction accuracy drops to 99.7% even when the number of users is increased to 100. The experiments for predicting the identity of the author of a message show that each author has a distinct communication style and word selection habit. The use of style-based feature sets also show that the receiver of a message and the time period in which the message is written are also predictable. The receiver of a message is predicted with 75.0% and 40.9% accuracy for the Receiver-2-25 and Receiver-10-26 test cases, respectively. The classification accuracies for the DayPeriod-2-34 and DayPeriod-4-37 test cases are 71.6% and 47.6%, respectively. Table 9 also summarizes the significance tests conducted on message-specific attributes.

The use of style-based feature sets perform equally with term-based feature sets when the number of classes is small. However, as the number of classes increases, the decrease in the prediction accuracy is more significant when using style-based feature sets than using term-based feature sets. The reason of this rapid decrease in the prediction accuracies is that the dimensionality of the style-based feature sets are much smaller than that of the term-based feature sets; and as the number of classes increases, all classifiers exhibit difficulties in differentiating the instances of different categories.

Contrary to the results of the experiments employed using the term-based feature sets, the receiver and day period of a message can only be predicted with lower accuracies using a style-based feature set. This interesting finding shows that the vocabulary use of a person is dependent on the target and the time of the message while the communication style is only dependent on the person writing that message.

For predicting the message-specific attributes, NB and SVM achieve best results among all classifiers. While both classifiers perform similarly for small number of classes, the experiments on the authors' identity show that as the number of classes increases SVM performs better than NB. The PRIM classifier performs the worst for all attributes for both term- and style based feature sets.

Table 8
Prediction accuracies of experiments conducted on message-specific attributes

Tag	Term-based feature set				Style-based feature set			
	k-NN	NB	PRIM	SVM	*k*-NN	NB	PRIM	SVM
Author-2-35	**100.0**	**100.0**	98.7	**100.0**	98.3	**99.7**	92.9	97.1
Author-10-26	98.7	**100.0**	74.4	99.9	84.0	89.1	51.7	**97.1**
Author-100-10	88.3	89.9	44.0	**99.7**	31.2	29.7	5.8	**78.9**
DayPeriod-2-34	66.2	**71.6**	48.8	60.7	59.9	**63.8**	54.3	59.6
DayPeriod-4-17	34.6	**47.6**	25.4	39.6	30.7	38.9	28.5	**41.6**
Receiver-2-35	60.0	**75.0**	51.6	67.0	58.5	**60.5**	53.7	53.4
Receiver-10-26	25.1	40.9	21.8	**41.1**	**12.4**	11.2	9.2	10.6

Table 9
Significance analysis conducted on message-specific attributes

	Term-based feature set		Style-based feature set	
Tag	*z*-Score	*p*-Value	*z*-Score	*p*-Value
Author-2-35	5.24	1.5e−7	4.79	1.2e−5
Author-10-26	13.37	7.1e−73	13.12	9.2e−69
Author-100-10	27.19	3.3e−318	24.16	2.5e−200
DayPeriod-2-34	3.30	1.9e−3	1.46	1.8e−1
DayPeriod-4-17	2.32	3.9e−3	1.80	7.6e−2
Receiver-2-35	2.39	2.0e−3	1.20	2.4e−1
Receiver-10-26	6.18	5.2e−9	1.42	3.6e−1

7. Concluding remarks

7.1. Discussions

In this paper, the predictability of various user- and message-specific attributes in electronic discourse is examined. Specifically, the word selection and message organization of chat users are investigated by conducting experiments over a large real-life chat dataset. Our observations show that many characteristics of chat users and messages can be predicted using their word selection and writing habits. The experiments point out that some attributes have recognizable traces on the linguistic preferences of an author. A possible alternative view to the chat mining problem is to examine how the linguistic traits of a person effect the writing style. In this section, we take this alternative view and discuss how a person's attributes affect his writing style.

Table 10 shows the set of most discriminative terms for different attributes. As chat conversations occur in a spontaneous environment, the use of slang words and misspellings is frequent. Two different users may write the same word quite differently. For example, the word "something" (spelled as "birsey" in Turkish with ASCII characters) is used in its syntactically correct form by the user "Andromeda" while "Paprika" uses a slang version ("bishiy" in Turkish with ASCII characters) of the same word in his messages. The receiver of a message also affects the word selection habits. Some users tend to receive messages containing more slang words than others. The vocabulary use is additionally affected from the period of the day. Our observations show that during the day hours, users tend to converse more politely, using apologetic words more frequently.

The user-specific attributes also affect the word selection habits. The most discriminative words of the users connected from the ".edu" domain contain more inquiries and imperatives. On the contrary, the users connected from the ".com" domain employ mostly responses and second person (formal) references. The users

Table 10
The most discriminating words for each attribute[a]

Attribute name	Example class	The most discriminating words
Author	Andromeda	byes (bye – slang), ok, birsey (something)
	Paprika	diil (nothing – misspelled), ehe (hah – slang)
		bishiy (something – misspelled)
	Taru	hmm (emoticon), dakika (minute), ha (hah!)
BirthYear	1979	dusunuyon (thinking – misspelled), ucuza (cheaply)
		acar (opens)
	1978	onemli (important), demek (then), git (go)
DayPeriod	Afternoon	kusura (fault), uzgunum (I'm sorry), lutfen (please)
	Evening	geceler (nights), hosca (finely), grad (graduate)
Domain	.edu	git (go), gelir (comes – 2nd person), saat (clock)
	.com	cikardin (you displace – 2nd person), muhabbet (chat)
		karsindaki (opposite)
Gender	male	abi (brother), olm (buddy – misspelled)
		lazim (required)
	female	ayyy (ohhh!), kocam (my husband)
		sevgilimin (my lover's)
Receiver	Celefin	olm (buddy – misspelled), falan (so)
		yaw (hey! – misspelled)
	Kebikec	hmm (a notification), seker (sugar), adam (man)
School	Ege University	Ege (a region), Bornova (a city in Agea region)
		Izmirde (in Izmir, a city in Agea region)
	Bilkent University	Bilkent (University Name), BCC (Bilkent Computer Center)
		Bilkentte (in Bilkent)
	METU University	ODTU (University Name in Turkish), METU (University name)
		yurtlar (dormitories)

[a] The discriminative power of each word is calculated using the χ^2 statistic.

of the ".com" domain tend to use shorter words than the users connected from other domains in their conversations. Another attribute that clearly affects the vocabulary of a user is gender. It is apparent that males tend to use more decisive, dominating sentences using words that can be considered as slang while female conversations involve more content-dependent words and emoticons (e.g., Ayyy!). These findings show similarities with the findings presented in (Zelenkauskatie & Herring, 2006). The most discriminative words for the classes of user's educational environment are mostly dominated by the regional terms. In Table 10, the most discriminating words of users from three universities in different regions are given. The vocabulary of the users contain many location-specific terms and is clearly affected by the location of the university and its facilities.

The stylistic analysis also provides interesting results. Each chat user expresses himself/herself using an almost-unique and identifiable set of linguistic preferences. The messages of three different users is examined in order to present their stylistic differences. The user named "Andromeda" employs smileys and average-length words more than others, while "Paprika" tend to converse using shorter messages, prevent using punctuation marks, smileys, and function words. The user "Taru" communicates with longer messages containing a large number of punctuation marks and function words. The time of a message also affects the style and vocabulary of a message. During the day hours, messages are generally shorter and contain less auxiliary elements such as smileys and punctuation marks, while during the night hours the messages tend to be longer containing many function words and punctuation marks.

The writing style shows variations between different domains. The users connected from the ".edu" domain have a smaller vocabulary and use punctuation marks and numerals frequently. On the contrary, the users of the ".com" domain have a larger vocabulary, use a small number of numerals, and write longer messages. The educational environment of a user is another factor that affects the writing style. The users from different universities prefer to use separate sets of smileys. The style of a person is also affected by his/her gender. In general, female users prefer longer and content bearing words. They also prefer shorter sentences than male users and omit the use of stopwords and punctuation marks. Long messages and use of short words are most discriminating stylistic characteristics of male users. The use of style-based feature sets prove to be more effective than the use of term-based feature sets for determining the birth year of an author. This result also shows that the age group of an individual is an important factor that affects the stylistic characteristics of a person's messages. The experiments conducted for determining the birth year attribute of a user show that younger users mostly have a smaller vocabulary. Additionally, as Radford and Connaway (2007) also pointed out, younger users prefer using smileys more than older users.

7.2. Conclusion

The result of this study show that personal and environmental characteristics have significant impact on ones' vocabulary use and writing style in peer-to-peer communications. In this paper, it is shown that by using the word selection patterns and stylistic preferences of chat users, it is possible to predict their sociolinguistic characteristics by employing classification techniques. It is also shown that external factors such as the time of a conversation and the recipient of a message has considerable effect on the vocabulary use and writing style of an author.

The dataset used in this work also has distinguishing properties. The spontaneous nature of chatting and point-to-point nature of the chat messages makes the chat dataset quite different from any literary writing. To the best of our knowledge, in this study, for the first time in literature, the authorship analysis techniques are applied to real-time online conversations.

We believe that the outcome of this work will prove to be beneficial for many application areas such as e-commerce and Internet security. For example, it is possible that companies supporting virtual reference services may use this method for gathering client profiles, determining a target population, and provide better and more customized service to these clients. With the growing use of Internet communication, spamming becomes a worldwide phenomenon. This application can also be used in the implementation of dynamic spam filters. Once the classifier is trained by a set of previously available spam messages, it may be possible to identify the structural properties of spam messages and detect them. The style-based approach presented in this paper may prove to be useful for this purpose. Another direct implication is the use of our work for ensuring security within virtual groups. In most messaging services, a user is not permitted to have more than one

account. Matching user profiles may prevent duplicate user accounts and can be used to detect the true source of malicious messages.

This work can be extended in several ways. First, our approach is tested using only one corpus. Application of our methods on different datasets will strengthen the findings of this paper. Applying our methods to other types of electronic discourse such as emails, IRC messages, and newsgroup messages may reveal similarities between different computer-mediated communication media. Second, this work has only been tested on Turkish documents. While the applied procedure seems to be independent of the language, the effectiveness and applicability to other languages remain untested. Additionally, such a work may provide clues on common and language-independent characteristics of electronic discourse. Third, this work relies on the supervised learning assumption. This means that the procedures described here are applicable only if a set of training samples is available. A framework based on unsupervised classification seems to be a natural extension of this work. In the unsupervised classification approach, the classifier generates a set of spectral classes without requiring any input. Information classes are assigned to these spectral classes afterwards with user intervention. Fourth, the problem can be modeled as a probabilistic information retrieval model. Using the procedure described in this paper, it may be possible to answer queries such as "find the documents that are predicted to be written during a certain period of time" or "find the documents that are possibly written by someone who has a PhD degree".

Acknowledgements

We thank the anonymous referees for their constructive comments which led to a significant improvement of our paper.

References

Argamon, S., Saric, M., & Stein, S. S. (2003). Style mining of electronic messages for multiple authorship discrimination: First results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 475–480). Washington, DC, USA.

Baayen, R. H., van Halteren, H., & Tweedie, F. J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*(3), 121–132.

Backer, E., & van Kranenburg, P. (2004). Musical style recognition – a quantitative approach. In *Proceedings of the conference on interdisciplinary musicology (CIM04)*. Graz, Austria.

Binongo, J. N. G., & Smith, M. W. A. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing, 11*(3), 121–131.

Burrows, J. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Cambazoglu, B. B., & Aykanat, C. (2005). *Harbinger machine learning toolkit manual*. Tech. Rep. BU-CE-0503, Bilkent University, Computer Engineering Department, Ankara.

Can, F., & Patton, J. M. (2004). Change of writing style with time. *Computers and Humanities, 38*(1), 61–82.

Corney, M. W. (2003). Analyzing E-mail text authorship for forensic purposes. M.S. thesis, Queensland University of Technology.

Corney, M., Anderson, A., Mohay, G., & Vel, D. O. (2001). Mining email content for author identification forensics. *SIGMOD Record Web Edition, 30*(4), 55–64.

Dickey, M. H., Burnett, G., Chudoba, K. M., & Kazmer, M. M. (2007). Do you read me? Perspective making and perspective taking in chat communities. *Journal of the Association for Information Systems, 8*(1), 47–70.

Elliot, W. E. Y., & Valenza, R. J. (1991). Was the Earl of Oxford the true Shakespeare? A computer aided analysis. *Notes and Queries, 236*, 501–506.

Foster, D. W. (2000). *Author unknown: On the trail of anonymous*. New York: Henry Holt.

Graham, N., Hirst, G., & Marthi, B. (2005). Segmenting documents by stylistic character. *Natural Language Engineering, 11*(4), 397–415.

Hakuta, K. (1991). Bilingualism as a gift. In *Stanford Center for Chicano research working paper series 33*.

Han, E., Karypis, G., & Kumar, V. (2001). Text categorization using weight adjusted *k*-nearest neighbor classification. In *Proceedings of the fifth Pacific-Asia conference on knowledge discovery and data mining* (pp. 53–65).

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variations in Weblogs. *Journal of Sociolinguistics, 10*(4), 439–459.

Herring, S. C., & Danet, B. (Eds.). (2007). *Multilingual Internet: Language, culture, and communication online*. New York: Oxford University Press.

Holmes, D. I. (1985). Analysis of literary style – A review. *Journal of the Royal Statistical Society, 148*(4), 328–341.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities, 28*(2), 87–106.

Holmes, D. I., & Forsyth, R. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing, 10*(2), 111–127.

Hota, S., Argamon, S., & Chung, R. (2006). *Gender in Shakespeare: Automatic stylistics gender classification using syntactic, lexical, and lemma features*. Chicago, IL: Chicago Colloquium on Digital Humanities and Computer Science.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of 10th European conference on machine learning (ECML-98)* (pp. 137–142). Heidelberg, Germany.

Jonsson, E. (1998). Electronic discourse-on speech and writing on the Internet. Retrieved June 24, 2006, from www.ludd.luth/jonsson/D-essay/index.html.

Juola, P., & Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing, 20*(1), 59–67.

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th international conference on computational linguistics* (Vol. 2, pp. 1071–1075). Kyoto, Japan.

Kessler, B., Nunberg, G., & Schutze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th annual meeting on association for computational linguistics* (pp. 32–38). Madrid, Spain.

Khmelev, D. V., & Tweedie, F. J. (2001). Using Markov chains for identification of writers. *Literary and Linguistic Computing, 16*(3), 299–307.

Kjell, B. (1994). Authorship attribution of text samples using neural networks and Bayesian classifiers. In *IEEE international conference on systems, man and cybernetics*. San Antonio, TX.

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401–412.

Krusl, I., & Spafford, E. H. (1997). Authorship analysis: Identifying the author of a program. *Computers and Security, 16*(3), 233–257.

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced data sets: One-sided sampling. In *Proceedings of the 14th international conference on machine learning* (pp. 179–186). Nashville, TN.

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2006). Chat mining for gender prediction. In *Proceedings of the fourth Biennial conference on advances in information sciences* (pp. 274–284). Izmir, Turkey.

Lam, W., Ruiz, M. E., & Srinivasan, P. (1999). Automatic text categorization and its applications to text retrieval. *IEEE Transactions on Knowledge and Data Engineering, 11*(6), 865–879.

Levitan, S., & Argamon, S. (2006). Fixing the federalist: Correcting results and evaluating editions for automated attribution. In *Digital humanities* (pp. 323–328). Paris, France.

Love, H. (2002). *Attributing authorship: An introduction*. Cambridge: Cambridge University Press.

McCallum A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*. Madison, WI.

Merriam, T., & Matthews, R. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing, 9*, 1–6.

Mosteller, F., & Wallace, D. L. (1964). *Inference and disputed authorship: The Federalist*. Reading: Addison-Wesley.

Patton, J. M., & Can, F. (2004). A stylometric analysis of Yasar Kemal's Ince Memed tetralogy. *Computers and the Humanities, 38*(4), 457–467.

Pollatschek, M., & Radday, Y. T. (1981). Vocabulary richness and concentration in Hebrew biblical literature. *Association for Literary and Linguistic Computing Bulletin, 8*, 217–231.

Radford, M. L. (2005). Encountering virtual users: A qualitative investigation of interpersonal communication in chat reference. *Journal of the American Society for Information Science and Technology, 57*(8), 1046–1059.

Radford, M. L., & Connaway, L. S. (2007). "Screenagers" and live chat reference: Living up to the promise. *Scan, 26*(1), 31–39.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities, 31*(4), 351–365.

Sabordo, M., Chai, S. Y., Berryman, M. J., & Abbott, D. (2005). Who wrote the "Letter to the Hebrews"?: Data mining for detection of text authorship. *Proceedings of the SPIE, 5649*, 513–524.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1–47.

Spafford, E. H., & Weeber, E. H. (1993). Software forensics: Can we track code to its authors? *Computers and Security, 12*(6), 585–595.

Stamatos, E., Fakotakis, N., & Kokkotakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics, 26*(4), 471–495.

Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology, 40*(2), 193–208.

Tsuboi, Y., & Matsumoto, Y. (2002). Authorship identification for heterogeneous documents. M.S. thesis, Nara Institute of Science and Technology.

Turell, M. T. (2004). Textual kidnapping revisited: The case of plagiarism in literary translation. *The International Journal of Speech, Language and the Law, 11*(1), 1–26.

Tweedie, F. J., Singh, S., & Holmes, D. I. (1996). Neural network applications in stylometry: The federalist papers. *Computers and the Humanities, 30*(1), 1–10.

Vel, D. O., Corney, M., Anderson, A., & Mohay, G. (2002). Language and gender author cohort analysis of e-mail for computer forensics. In *Second digital forensics research workshop*. Syracuse, USA.

Walther, J. B., Bunz, U., & Bazarova N. N. (2005). Rules of virtual groups. In *Proceedings of the 38th annual Hawaii international conference on system sciences*. Big Island, HI.

Wikipedia. (2007). Emoticon. Retrieved October 23, 2007, from <http://en.wikipedia.org/wiki/Emoticon>.

Wilcoxson, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80–83.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the fourteenth international conference on machine learning* (pp. 412–420). Nashville, TN, USA.

Zelenkauskatie, A., & Herring, S. C. (2006). Gender encoding of typographical elements in Lithuanian and Croatian IRC. In *Proceedings of cultural attitudes towards technology and communication* (pp. 474–489). Tartu, Estonia.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style and classification techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378–393.