IET Research Journals

This is the uncorrected author proof; copyright rests with IET Research Journals. Please cite this work as Kuzu, R.S., A.A. Salah, "Chat Biometrics," IET Biometrics, accepted for publication.

IET Journals
The Institution of Engineering and Technology

CHAT BIOMETRICS

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Rıdvan Salih Kuzu[1,2*], *Albert Ali Salah*[3,4]

[1] *Department of System and Control Engineering, Boğaziçi University, 34342-Istanbul, TURKEY*
[2] *Department of Applied Electronics, Roma Tre University, 00146-Roma, ITALY*
[3] *Department of Computer Engineering, Boğaziçi University, 34342-Istanbul, TURKEY*
[4] *Future Value Generation Research Center, Nagoya University, Nagoya, JAPAN*
* *E-mail: ridvansalih.kuzu@uniroma3.it*

Abstract: On-line social platforms implement moderation mechanisms to filter out unwanted content and to take action against possible cases of verbal aggression and abuse, sexual harassment, and such. In this study we investigate chat biometrics, the identification of users from their verbal behaviour on a social platform. The typical application scenarios are the re-identification of banned users, returning under different identities, and aggressors operating through multiple fake accounts. We propose a novel processing pipeline, and contrast the problem with the authorship recognition problem, which is relatively well-studied in the literature. We evaluate our proposed approach on a large corpus of multiparty chat records in Turkish, which we have previously collected from a multiplayer game environment. We also introduce a new corpus in this study, collected from a well-known Turkish social platform called Ekşisözlük, in order to test the robustness of the system across domain changes, as well as on Portuguese and English news datasets to test it on different languages. We evaluate both instance-based and profile-based approaches, and provide detailed analyses with regards to the required amount of text to identify a person reliably.

1 Introduction

Computer-mediated communication with text messages has become very prevalent in the Internet era. In addition, instant messaging applications on mobile devices have received widespread attention, thousands of multiplayer online games and dedicated platforms provide chat facilities. These Internet-based services constantly generate large amounts of text data. In this study, we explore the degree to which a person can be identified from chat communications, which we call "chat biometrics".

While content and style of text messages depend on many factors, it may be possible to match an unsuspected person by using a pre-collected corpus. It may also be possible to deduce gender, age, and ethnicity, based on the specific words and forms used during chat, and based on particular mistakes. The analysis of writing style was investigated in the context of "authorship recognition," which seeks the identification of the author of a text among a set of candidate authors, whose texts are available for supervised classifier training. The electronic chat domain is significantly different from the literary text domain. These differences are particularly prominent in word and character frequencies, use of punctuation marks, intentional and unintentional misspellings, vocabulary usage, sentence length, and the particular ordering of words. The increased freedom in the usage of language, coupled with (typically) much more limited vocabulary makes chat biometrics an interesting challenge.

We are motivated in this work by moderation mechanisms implemented for online social platforms, for which it is essential to filter out unwanted content and to take action against possible cases of verbal aggression and abuse, sexual harassment, and such [1]. Automatic evaluation of aggression cases is typically performed via user profiling, and the textual content of interaction is not processed [1, 2]. In this work, we propose a text-based system for monitoring the platform for repeated offenders. We mainly use a corpus acquired from the chat interface of a multiplayer online game, together with meta-data concerning more than a thousand complaints filed by players. In such games, offending users who are blocked by administrators for various reasons (such as cheating, foul language, hate speech, abusive behaviours) may return to the game using an impostor account. Finding these matching accounts is a very hard problem to tackle manually. Game communities spend resources to preserve a user friendly gaming environment, which includes offending players. Reducing the number of suspects might be very useful, even if finding the real offender is difficult.

We investigate the rate of success in identifying malicious users in a multi-participant chat environment by means of extracting relevant features and recognition techniques. In our approach, we apply and compare several methods to match users to a gallery by their chat records. We test our approach with documents that have Turkish chat content, which brings additional challenges due to the agglutinative nature of the language (i.e. many postfixes are applied on word roots). We discuss these briefly, and assess different strategies for dealing with them. The proposed approach is applicable to other languages.

The work introduced here is closely related to "chat mining" [3], where text from online social platforms are mined for specific purposes, such as identifying the unknown author of a post among suspects. In particular, we study the following questions:

- What is the effectiveness of existing author identification methods for attributing authorship of a set of chat texts to one person among a closed set of suspects?
- What are the influential and effectual features of chat messages for the purposes of chat biometrics?
- How much text is required to extract an accurate author profile?
- How can we improve the author recognition performance?

We extend our earlier work [4] in a number of ways: i) measuring effects of dictionary size on performance, ii) comparing weighting schemes to be used in authorship analysis, iii) suggesting a novel recognition pipeline which gives better results than baseline and our previous study, iv) comprehensive literature comparison with concurrent studies in chat biometrics, and v) measuring the effects of domain changes with regards to intra-language and inter-language variations by experiments on three additional databases.

In Section 2 we provide a survey for this relatively new domain. Section 3 describes the proposed methodologies for identifying people from their chat messages. We introduce a novel database in Turkish and three existing databases (in Turkish, English, and Portuguese) used in experimental evaluations of this study in Section 4. Section 5 provides the experimental results with discussions. Section 6 highlights the main findings.

2 Related Work

2.1 Authorship Analysis

Authorship analysis aims to identify individuals by the statistical properties and characteristics of their language use. To distinguish text written by different authors, textual features, as well as machine learning techniques can be employed. Gray *et al*. identified several approaches of authorship analysis that can be applied to software forensics [5]. Based on their definitions and other related studies, authorship analysis can be grouped into five major categories, as summarised in Table 1.

Table 1 *Types of authorship analysis in summary*

Category	Description	Label
Authorship Recognition	Uses a training set of different authors' writings to determine the likelihood of authorship on a new piece of writing	AR
Authorship Verification	Uses a set of documents by an author, determines whether a new document is also by that author or not.	AV
Authorship Profiling	Determines an author profile by summarising features obtained from the works of the author.	AP
Authorship Discrimination	Given a document or a corpus, decides whether it is written by a single author or by multiple authors without actually identifying who they are.	AD
Author Intent Determination	Seeks certain intentionally produced properties of a given document or corpus, including style.	ID

Authorship attribution means finding the author of a document. To achieve this purpose, one compares a query text with a model of the candidate author and determines the likelihood of the model for the query. One of the early attempts on authorship analysis was performed by Mendenhall [6]. Mendenhall looked into word lengths, comparing Dickens, Thackeray, and Mill, and decided that a hundred thousand words were enough to determine a signature for an author. The most curious series of works in this area were realised by Mosteller and Wallace in the 60s [7]. In their study of the authorship attribution on 146 political essays (known as the Federalist Papers), a Bayesian approach was applied on the frequencies of a small set of common words, and promising results were obtained. Their conclusions have shown that historical information can be obtained through such text analysis [8, 9].

Authorship attribution can be considered in two different categories: i) *authorship recognition*, and ii) *authorship verification*:

In the recognition mode, a script from an unknown author is compared with all the authors' textual records for a match. It is a one-to-many comparison to detect the identity of an author, and this attempt should fail if the author is not enrolled in the database before [10].

In the verification mode, the system receives a text together with a claimed identity, and compares the textual data with he previously stored scripts of the particular author. Typically, a similarity will be computed based on extracted features, and a threshold value will be checked to determine whether the query text was written by the author in question [11].

Authorship profiling or characterisation aims to find sociolinguistic patterns in the writings of an author to determine certain attributes. Some attributes previously examined in the literature are gender, educational and cultural background, and language familiarity.

Authorship discrimination aims to determine if a document or a corpus is written by a single individual, or by multiple authors. Inter-subject and intra-subject variability of a text are computed in such problems to determine the validity of the claim that two texts belong to different authors without regarding their identities. Many studies in this field are also related with plagiarism detection. Plagiarism is the partial or complete replication of a piece of work, and plagiarism detection is used for investigating suspicious documents against potential original documents [12].

Author intent determination, as initially defined in the code domain, aimed to detect intentionally malicious code [5]. In a biometric setting, it can refer to the detection of any stylistic or content-related property of a document produced by the author.

These problems can be adopted to a chat setting, where the system typically has access to some additional profile information about the user, but has no guarantee of the correctness of this information. While profile features can be beneficial in authorship analysis, we do not tackle them in this work. The next section describes the most commonly used features in the literature.

2.2 Feature Types

The state-of-art approaches in authorship analysis depend on stylometric features, which can be divided into six major categories: i) word, ii) character, iii) syntactic, iv) structural, v) content specific, and vi) semantic features, respectively. A brief description and the relative discrimination capability of each type of feature are given next.

Word features are used to learn about the preferred use of words of an individual. Pioneering efforts on attributing authorship were based on trivial measures like counts of sentence length and word lengths [8]. The use of such features illustrates the tendency of an individual to use particular words or phrases. They can be easily applied to any language and any textual database without the need for additional tools, apart from a tokeniser, which divides the texts into tokens (e.g., words, characters).

Character features include frequency of individual symbols in the alphabet, total number of upper/lower case letters, distribution of capital letters used in the beginning of sentences, average number of characters per word, and average number of characters per sentence, etc. [9]. In this domain, extracting frequency distribution of character n-grams (i.e. strings of length n) is a more comprehensive and also computationally simpler procedure. Additionally, this method is capable of catching style nuances along with lexical information, contextual marks, as well as punctuation and capitalisation preferences. Representation of text in the n-gram domain is more robust to noise, compared to word representations. This is an important point especially for the chat domain that we tackle in this paper, as chat messages are very noisy, and a single word can have many different alternatives.

Syntactic features, such as part-of-speech (POS) tags, chunks, sentence and phrase structure, are sentence-level features for capturing an author's writing style. Sentence organisation can be a cue to detect authorship. While function words do not contribute much to semantics, they describe relationships between content words, and their distribution and specific usage can be informative [13].

Structural features depend on distinctive habits of people while organising a document, including length of paragraphs and visual layout. Online documents are particularly rich in such features, as their structuring is much easier than printed text. Structural features were first suggested by de Vel *et al*. for e-mail authorship attribution and led to high identification performance [14].

Content-specific features are words or features that are indicative of a particular activity or social setting. To illustrate, messages aiming cybercrime activities like fraudulent sale offers, spamming and phishing frequently have phrases containing slang or street words [9]. Generally, features extracted for a specific domain cannot be directly applied to other domains.

Semantic features are used for determining semantic resemblances among words and phrases with the assistance of linguistic analysis. Synonyms, antonyms, hyponyms and hypernyms of the words can be explored, and existing NLP tools are very useful for such analysis. However, extracting complex semantic features is difficult, and for many languages (including Turkish), there are few adequate tools [8]. Based on these feature categories, some of the most commonly used feature types to represent a text for authorship analysis are listed in Table 2.

If we consider simplicity and language independence as primary factors, lexical features are expected to perform better than other features. Especially, the character *n*-gram representation has been used as one of the most effective measures of authorship attribution [13, 15]. If authors tend to use similar patterns in their writings, this would imply that syntactic and semantic features may lead to superior results. On the other hand, language-specific NLP tools like part-of-speech taggers, stemmers, spell checkers etc. are needed to exploit these features.

Text analysis can be significantly different on agglutinative languages such as Hungarian, Finnish and Turkish. Such languages have complex word structures, which are formed by stringing together morphemes without changing them in spelling or phonetics. In Turkish, for example, the word *ev-ler-iniz-den* would translate to *house-plural-your-from*, i.e. *from your houses*. This causes difficulties in stemming and in syntactic analysis, particularly for noisy text obtained from social media. Moreover, transposed sentence structure is very common and word order in a sentence is very flexible, without changing the meaning. For instance, a sentence with three words like *ben eve geldim* (*I came home*) can be expressed with six different word orders in Turkish. Hence, even if the same words are used, ordering habits can give some hints about the author. Consequently, one of the the simplest and effective ways of representing an author is using *n*-grams on such languages.

There are some prior studies on authorship attribution in Turkish. Tufan *et al*. used style marker features on a gallery of 20 authors [16]. Amasyali *et al*. categorised texts in terms of author (18-class), genre (3-class) and gender (2-class) by using stylistic and *n*-gram features [17, 18]. In these studies, success rates around 80% were reported with a Naive Bayes classifier. Both studies used newspaper articles, but neither the gallery size, nor the domain characteristics are representative for chat biometrics. Furthermore, they use language-specific features, whereas our approach is applicable to different languages.

2.3 Analysis Techniques

A typical authorship recognition problem contains a set of text samples for candidate authors, and query text samples from unknown authors. Each sample should be attributed to a candidate author. Identification approaches can be distinguished as profile-based and instance-based, according to whether the set of text samples for each author is treated individually, or cumulatively [8].

Concatenating training texts per author in one single text file is known as the *profile-based approach* (PBA). This large single file is used to extract properties of the author's style. A text sample from an unknown author is compared with each author profile, and a suitable distance measure is used to find the most likely author. In this approach, features related with the variety of texts in the training corpus are not taken into consideration.

Instance-based approach (IBA), on the other hand, considers each text sample independently, hence the differences in the training texts by the same author are not neglected. Both approaches have their own advantages, but if text documents are very concise and limited, concatenation of the text (as in profile-based approaches) may help to create a sufficiently long document for capturing the author's style [19]. Instance-based approaches are believed to be more effective when sufficient amount of text per author is available. However, Potha and Stamatatos [20] reported the best results on the PAN-2013 Authorship Analysis Competition with a profile-based approach.

Performance in this domain also depends on pre-processing techniques, document set sizes, weighting schemes, language characteristics, and feature sets. In terms of used features, character *n*-grams, word tokens, distribution-based similarity features are typically preferred. Some common identification and attribution approaches in terms of feature extraction and matching methods are summarised in Table 3. For a recent survey of the broader field of authorship attribution, see [21]. More recently, convolutional neural networks are tested for this problem, using bigrams and word embeddings as features [22–25].

Table 2 *Commonly used features for authorship analysis*

Word Features (WoF)	**Character Features (ChF)**	**Structural Features (StF)**
-total # of words -total # of unique words -ratio of short words -mean word length -ratio of distinct words -# of hapax legomena -# of hapax dislegomena -word *n*-grams -skip-grams -word frequencies -# of words of each length -vocabulary richness	-total # of characters -ratio of alphabetic chars. -ratio of upper case letters -ratio of digit characters -ratio of white space chars. -ratio of tab space chars. -ratio of special chars. -ratio of emoticons -ratio of char. repetition -character *n*-grams -vowel combination -compression methods	-# of sentences -# of paragraphs -# of quoted content -# of lines -# of characters per paragraph -# of words per paragraph -# of sentences per paragraph -farewells -greetings -indentations -signature
Syntactic Features (SyF)	**Content Specific Features (CsF)**	**Semantic Features (SeF)**
-freq. of function words -freq. of punctuation marks -part of speech (POS) tags -total # of line -total # of sentences -ratio of spelling errors	-# of stop words -# of abbreviations -# of keywords -gender/age based words -slang words -writing speed -turn duration (for chat)	-synonyms of words -hypernyms of words -hyponyms of words -semantic dependency graphs -latent semantic analysis -systemic functional grammar -discourse features

2.4 Chat Biometrics

The bulk of authorship analysis approaches in the literature focus on the English language, and there are a few important studies related to chat biometrics on texts in English. Inches *et al*. [42] used two different internet relay chat (IRC) datasets containing homogeneous and heterogeneous topics separately. Traditional chi-squared distance and KL divergence were used to determine the similarity between the author profiles. The study achieved up to 61% accuracy on heterogeneous chat records. Layton *et al*. [54] used IRC records of 50 users (50 chat messages for each). The re-centred local profile (RLP) method was used for identification. Using an ensemble classification scheme, where each classification was weighted by the ratio between the distances to the second closest and closest authors, 55% accuracy was achieved.

Roffo *et al*. [44] proposed to adopt features inspired by conversation analysis (in particular for turn-taking), as well as to extract the features from separate turns rather than from entire conversations. The corpus used for their study contains 312 dyadic Italian chat conversations, collected with Skype over a time span of 5 months. They report a recognition rate of 76.9% on a total of 78 subjects. Their approach does not take the actual content into consideration, but focuses on different features for analysis, such as character writing speed, total chat time, or other features of temporal nature. These features are typically not stored by software that handle chat records. Subsequently, chat based biometrics can be extended to behavioral biometrics, only if special software requirements are met. In this paper, we focus on the much more common case where such information is discarded.

In a related work to ours, a chat mining framework was tested on a Turkish dataset containing peer-to-peer text messages [3]. This work is one of the most exhaustive efforts on chat biometrics in Turkish, and while it does not cover multiparty chat, it established that context plays a significant role on vocabulary use and writing style in peer-to-peer communications. The authors reported that term-based features achieved better results compared to style-based features on a 100-author problem.

The problem of adversarial attacks has not been explored in the chat domain, but for literary texts, modifying a text to resemble some specific author will fool most automatic detection approaches [55]. While the imitation of style seems difficult in the chat domain, as the "style" is more erratic than a literary text, we investigated how it would be possible to mimic another user, given some sample of the user's chat records. It appears that the most important cues are the usage of the most frequently words (such as greetings and congratulations) and the amount, type, and frequency of emoticons. We

Table 3 *Summary of studies in authorship analysis*

Previous Studies	Category				Features						Techniques				# of
	AR	AP	AV	AD	WoF	ChF	SyF	StF	CsF	SeF	PBA	IBA	Detail	Language	Subjects
Stamatatos et al.'00 [11]	✓		✓		✓		✓		✓			✓	RM,DA	Greek	10
De Vel et al.'01 [14]	✓				✓	✓	✓	✓				✓	SVM	English	3
Kešelj et al.'03 [26]	✓					✓					✓		CNG,PD	Multiple	10
Clough'03 [27]				✓	✓		✓					✓	SM	English	9
Amasyalı & Diri'06 [17]	✓	✓				✓						✓	NB,SVM,RF,DT	Turkish	18
Zhao et al.'06 [28]	✓						✓				✓		KLD	English	7
Zheng et al.'06 [9]	✓				✓	✓	✓	✓	✓			✓	DT,NN,SVM	Multiple	20
Juola'06 [13]	✓					✓					✓		CE	Multiple	13
Sanderson & Guenter'06 [29]			✓		✓	✓					✓		MC	English	50
McCarty et. al.'06 [30]	✓				✓	✓	✓			✓		✓	Coh-Metrix	English	3
Frantzeskou et al.'07 [31]	✓					✓					✓		SCAP	C++ / Java	8
Tufan & Görür'07 [16]	✓				✓		✓					✓	NB	Turkish	20
Meyer zu Eissen et al.'07 [12]				✓	✓		✓	✓				✓	DA,SVM	English	N/A
Estival et al.'07 [32]		✓			✓	✓	✓	✓	✓			✓	SVM,KNN,Bagging	Multiple	1,033
Küçükyılmaz et al.'08 [3]	✓	✓			✓	✓	✓		✓			✓	KNN,NB,SVM,PRIM	Turkish	100
Argamon et al.'09 [33]		✓			✓		✓					✓	BMR	Multiple	19,320
Koppel et al.'11 [34]	✓					✓					✓		SM	Multiple	10,000
Solorio et al.'11 [35]	✓				✓	✓	✓	✓				✓	SVM	English	100
Escalante et al.'11 [36]	✓					✓				✓		✓	LOWBOW,SVM	English	10
Oliveira et al.'12 [37]	✓					✓					✓		NCD	Portuguese	100
Layton et al.'12 [38]	✓				✓	✓	✓	✓			✓		RLP,PD	Multiple	13
Savoy'12 [39]	✓				✓		✓				✓		Z-Score	Multiple	20
Cristani et al.'12 [40]	✓				✓	✓	✓		✓		✓		BD, ED	Italian	77
Seidman'13 [41]			✓		✓	✓	✓				✓		GI	Multiple	20
Inches et al.'13 [42]	✓				✓				✓		✓		KLD, Chi-Square	English	1,502
Monaco et al.'13 [43]			✓		✓	✓	✓					✓	KNN,ED	English	30
Roffo et al.'13 [44]	✓		✓		✓	✓	✓		✓		✓		RKHS	Italian	78
Brocardo et al.'13 [45]			✓			✓					✓		SM	English	87
Iqbal et al.'13 [46]	✓	✓			✓	✓	✓	✓	✓			✓	K-Means,EM	English	158
Schwartz et al.'13 [47]	✓				✓	✓						✓	SVM	English	1,000
Mikros et al.'13 [48]	✓				✓	✓					✓	✓	SVM	Greek	10
Portha & Stamatatos'14 [20]			✓		✓						✓		CNG,SCAP	Multiple	35
Qian et al.'14 [49]	✓				✓	✓	✓					✓	SVM,RM	English	62
Seroussi et al.'14 [50]	✓				✓				✓	✓		✓	DADT-P,SVM	English	22,116
Segarra et al.'15 [51]	✓	✓		✓			✓				✓		WAN	English	10
Overdorf & Greenstadt'16 [52]	✓				✓	✓	✓					✓	RM,ADF	English	50
Ruder et al.'16 [22]	✓				✓	✓						✓	CNN,SVM,SCAP,LDAH-S	English	62
Rocha et al.'17 [21]			✓		✓	✓	✓				✓	✓	SVM,RF,SCAP	English	50
Stamatatos'17 [53]	✓		✓		✓	✓						✓	DV,SVM	Multiple	13
Wang et al.'17 [23]	✓					✓				✓		✓	CNN,SVM	English	62
Sari et al.'17 [24]	✓				✓	✓						✓	CNN	English	62
Shrestha et al.'17 [25]	✓					✓						✓	CNN	English	50

Abbreviation List of Techniques used in Literature

ADF	Augmented Doppelgänger Finder	**ED**	Euclidean Distance	**NN**	Neural Network
BD	Bhattacharya Distance	**GI**	General Impostor	**PD**	Profile Dissimilarity
BMR	Bayesian Multinomial Regression	**KLD**	Kullback-Leibler Distance	**PRIM**	Patient Rule Induction Method
CE	Cross Entropy	**KNN**	K-Nearest Neighbour	**RF**	Random Forest
CNG	Common N-Grams	**LDAH-S**	LDA Hellinger Single-Document	**RKHS**	Reproducing Kernel Hilbert Spaces
CNN	Convolutional Neural Network	**LOWBOW**	Locally-weighted bag of words	**RLP**	Recentered Local Profile
DA	Discriminant Analysis	**MC**	Markov Chains	**RM**	Regression Model
DADT-P	Probabilistic Attribution with Author-Document Topic Model	**MLP**	Multilayer Perceptron	**SCAP**	Source code author profiling
		NB	Naive Bayes	**SM**	Similarity Measure
DT	Decision Trees	**NCD**	Normalised Compression Distance	**SVM**	Support Vector Machine
DV	Distorted View			**WAN**	Word Adjacency Networks

suspect that a deliberate attempt at changing the style will easily succeed, but the typical use cases we deal with (i.e. detecting an abuser for repeated offence) makes this scenario unlikely.

3 Methodology

We propose two separate approaches for instance-based and profile-based author attribution. We summarise these briefly, before giving detailed explanations about sub-components.

Our instance-based authorship attribution model, illustrated in Figure 1, can be summarised as follows. The documents of a given author are concatenated into groups randomly, where the group size is a parameter to be determined empirically. If an author has 1,000 documents, and the group size is set to 20, then the author will have 50 enriched documents after concatenation, and the n-gram features will be extracted from these concatenated documents. This parameter obviously depends on the application and the average length of individual documents, and the minimum amount of text that can identify its author for that particular domain [56]. We preferred counting documents instead of words, since the amount of text per chat session and the number of words typical for a chat entry are stylistic features (absent for instance for Twitter posts).

In this work, character n-grams are preferred over word n-grams, as they are more suitable for shorter texts, and avoid the necessity of using complex NLP toolchains [26] that at the moment do not work well with highly noisy chat data [57]. A dictionary subset is extracted after ranking by using one of the mentioned feature selection methods in Section 3.1. The documents of the authors are represented with a vector space model, where columns are documents, and the rows are terms of the dictionary subset, as explained in Section 3.2. Global and local feature weighting schemes are applied on the vector space model after L_2 normalisation of document vectors. The weighted vector space model of author documents is transformed into a sub-space by using latent semantic analysis, and a multi-class extreme learning machine is trained with the transformed vector space model. When a text or group of texts from an unknown author are given, all the steps in the attribution model are applied and the

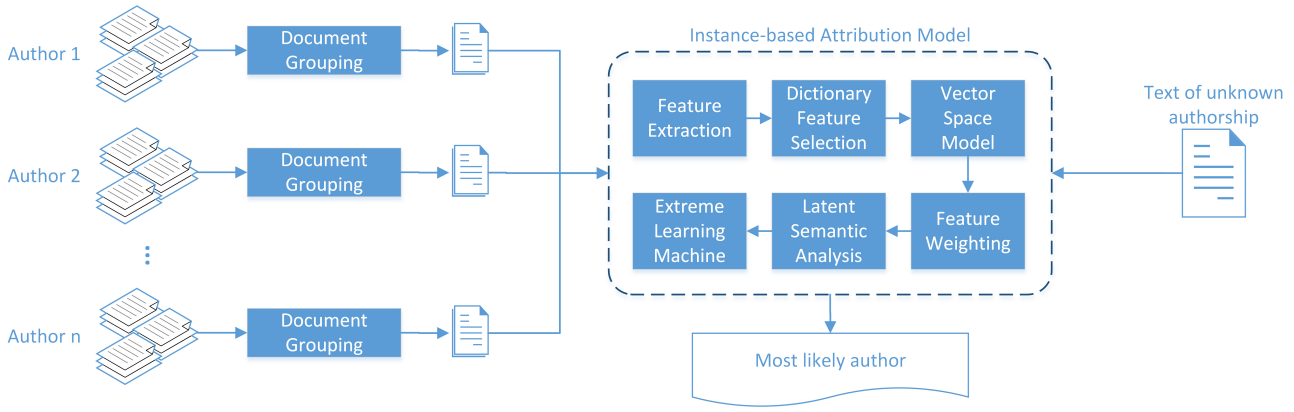


Fig. 1. ***Pipeline of the proposed instance-based approach.***

extreme learning machine predicts the most likely author of the query text.

For profile-based author attribution (illustrated in Figure 2), each author has only one concatenated document. In order to suppress noise in textual data and to represent author profiles in a compact manner, all profiles are transformed into a subspace with principal component analysis (PCA). For a given query, the profile that gives the minimum dissimilarity score is selected as the most likely author. In this study, cosine dissimilarity is the preferred distance measure.

3.1 Dictionary Feature Selection

Features extracted from character or word frequencies have generally high dimensionality. Especially, representation of a text in n-grams with $n > 2$ requires thousands of features. On the other hand, n-grams with higher n degrees do not only provide lexical information, but also provide clues about syntactic behaviours of an author. For that reason, there should be a trade-off between dictionary size and expressivity, which is domain specific. Furthermore, what part of the dictionary should be deemed relevant could also be a domain-specific question. For that reason, it is important to check whether high-frequency words or more discriminative words should be prioritised.

The extraction of a sub-dictionary (or equivalently, determining the cut-off frequency) can be handled with various approaches:

1. *Global Frequent Ranking:* All the features extracted from the dataset are ordered according to descending frequency, and the top k unique features are selected to represent the sub-dictionary.
2. *Local Frequent Ranking:* Features of each author are ranked in descending order separately. After that, the sub-dictionary of each author is determined by their own top k unique features. An example of such a ranking is given in the SCAP method [31].
3. *Local Distinctive Ranking:* Similar to local frequent ranking, each author is ranked with their own distinctive features in a descending order. Then, top k features are used to represent each author separately. Re-centring local profiles of each author according to global dictionary features is an example of such ranking [38].

3.2 Vector Space Representation

The vector space model (VSM) is a prominent approach in natural language processing applications [58]. In VSM, textual data can be represented as a vector of terms (bytes, characters, words, etc.). Based on this, assume there are n unique authors in the corpus $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}, ..\boldsymbol{a_i}.., \boldsymbol{a_n}]$, where each author $\boldsymbol{a_i} \in \boldsymbol{A}$ has varying number of documents totalling N. The profile of an author can be represented by the documents of the author, as $\boldsymbol{a_i} = [\boldsymbol{d_1^i}, \boldsymbol{d_2^i}, ...\boldsymbol{d_j^i}..., \boldsymbol{d_{n_i}^i}]$, where each $\boldsymbol{d_j^i}$ is one document of author $\boldsymbol{a_i}$ and n_i is the number of documents belonging to the author $\boldsymbol{a_i}$. In this representation, each document is represented as a fixed-size vector of frequencies in the term space, i.e. $\boldsymbol{d} = [f_{(t_1)}, ..., f_{(t_M)}]^T$, where M is the term (feature) set size, $t_i \in \boldsymbol{T}$ represents the term dictionary (or language profile), and $f_{(t_i)}$ is the frequency of term t_i in a given document $\boldsymbol{d}$. Then, the VSM is an $M \times N$ matrix composed of vector representations of all the documents in the author corpus:

$$\boldsymbol{A_{MxN}} = \begin{bmatrix} f_{d_1(t_1)} & \cdots & f_{d_N(t_1)} \\ \vdots & \ddots & \vdots \\ f_{d_1(t_M)} & \cdots & f_{d_N(t_M)} \end{bmatrix} \quad (1)$$

where $f_{d_j(t_i)}$ is the frequency of term i in document j. In this representation, each document corresponds to a column of term frequencies, and such frequency based representation disregards the order of terms in the document.

In our study, each chat session counts as a single document, but they are sometimes too short (e.g. 1-2 words or emoticons). To ensure representative profiles, chat entries of each author are concatenated to form sufficiently long documents.

3.3 Feature Weighting Schemes

Different terms (words, phrases, character combinations, or any other indexing units to identify the contents of a text) may have different importance for chat biometrics. Term weighting approaches can highlight distinctive features by assigning appropriate weights to the terms. Weighting schemes are based on two fundamental principles according to how they are used on VSM:

1. *Local Weighting Scheme:* If a term is used more frequently than others in a text or by an author, the term should have more importance than others.
2. *Global Weighting Scheme:* If a term is used commonly in different texts or by various authors, it is less distinctive than infrequently used terms. Hence, its weight or importance should be reduced.

Based on these basic principles, the weight of each term $t_i \in \boldsymbol{T}$ for the corpus $\boldsymbol{A}$ can be found in VSM:

$$W_{i,j} = G_i.L_{i,j}, \quad (2)$$

where $\boldsymbol{G_{(m\times 1)}}$ is a global weighting scheme for $t_i \in \boldsymbol{T}$ over all $\boldsymbol{d_j}$'s, and $\boldsymbol{L_{(m\times n)}}$ is a local weighting scheme for each $t_i \in \boldsymbol{d_j}$.

Some common global and local weighting schemes are summarised in Table 4. The global schemes typically involve summations over all documents or authors. In the binary schemes, the presence of a term results in a value of 1, and its absence a value of 0, regardless of the frequency of the term. In text categorisation and authorship identification, term frequency - inverse document frequency (TF-IDF) and its derivatives are the most widespread methods preferred for describing term-document weights in VSM,

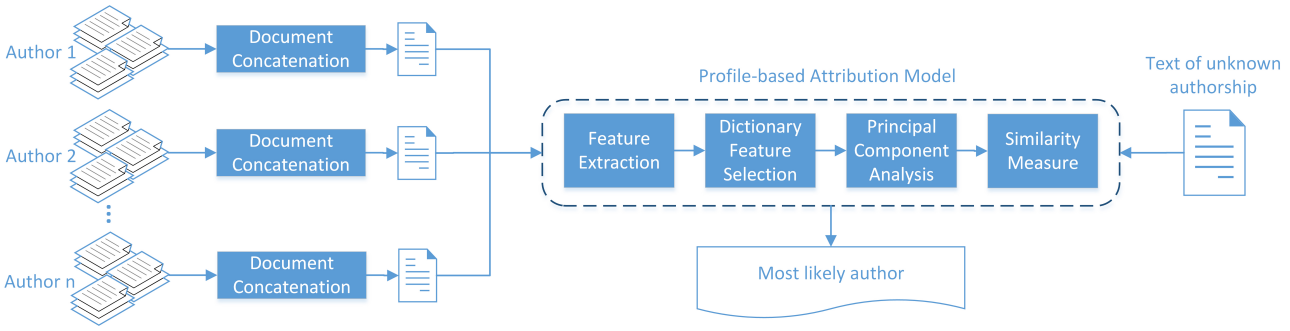


Fig. 2. ***Pipeline of the proposed profile-based approach.***

Table 4 Common weighting schemes used in the literature, where $f_{i,j}$ is the frequency of term t_i in document d_j.

Global Weighting Schemes		Local Weighting Schemes	
Scheme Formula	**Denotation**	**Scheme Formula**	**Denotation**
$\mathrm{G}_i \in \{1, 0\}$	binary	$\mathrm{L}_{i,j} \in \{1, 0\}$	binary
$\mathrm{G}_i = 1/\sqrt{\sum_j f_{i,j}^2}$	normal	$\mathrm{L}_{i,j} = f_{i,j}$	term frequency
$\mathrm{G}_i = n/\lvert d \in D : t \in d\rvert$	inverse document frequency	$\mathrm{L}_{i,j} = \log(1 + f_{i,j})$	logarithmic term frequency
$\mathrm{G}_i = 1 + \sum_j P(f_{i,j}) \log P(f_{i,j}) / \log n$	entropy	$\mathrm{L}_{i,j} = f_{i,j}/max_i(f_{i,j})$	augnorm
$\mathrm{G}_i = \sum_j f_{i,j}/\lvert d \in D : t \in d\rvert$	global frequency - IDF	$\mathrm{L}_{i,j} \in \{1 + \log f_{i,j}, 0\}$	sub-linear term frequency

because they are easy to implement, and work well with information retrieval tasks [59]. Although TF-IDF is reported to be useful for authorship recognition and text classification [54, 60, 61], its usefulness depends on the domain. Consequently, we compare multiple weighting approaches in this paper on different datasets.

3.4 *Subspace Projection*

Latent semantic analysis (LSA) and principal component analysis (PCA) are two eigenvalue methods typically used for dimensionality reduction of high-dimensional text based datasets. LSA is computed on the term-document matrix, while PCA is calculated on the covariance matrix. Based on this difference, PCA is preferred to transform author profiles in the profile-based method, as each author is represented with a single document. On the other hand, LSA is used for recovering correlations between terms and documents in the instance-based model, as such correlations can be ignored by VSM while representing documents.

3.4.1 Latent Semantic Analysis: LSA is an automated mathematical/statistical approach for extracting and deducing relationship among expected contextual usage of terms, words and phrases in passages of texts [62]. It is based on singular value decomposition (SVD) of vector space model such that:

$$\boldsymbol{W} \approx \tilde{\boldsymbol{W}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{3}$$

where $\boldsymbol{W}$ is the weighted term-document matrix, $\boldsymbol{U}_{(MxR)}$ is the matrix of left singular vectors $\boldsymbol{u_i}$'s $(1 \leq i \leq M)$,$\boldsymbol{\Sigma}_{RxR}$ is the diagonal of singular values, and $\boldsymbol{V}_{(NxR)}$ is the matrix of right singular vectors $\boldsymbol{v_j}$'s $(1 \leq i \leq N)$. In other words, $\boldsymbol{u_i}$ and $\boldsymbol{v_j}$ are the projections of $t_i \in \boldsymbol{T}$, and $d_j \in \boldsymbol{A}$ respectively from the initial vector space model onto semantic representation domain [60].

The decomposition provides two different advantages: Firstly it eliminates sparsity by preserving significant elements of $\boldsymbol{W}$, secondly it makes possible to truncate left and right singular vectors depending on the size of R.

Let the j^{th} weighted document in the term-document matrix $\boldsymbol{W}$ be $\boldsymbol{d}_j^w$. Depending on Eq. 3:

$$\boldsymbol{d}_j^w = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{v}_j^\top, \tag{4}$$

$$\boldsymbol{v}_j = (\boldsymbol{d}_j^w)^\top \boldsymbol{U}\boldsymbol{\Sigma}^{-1} \tag{5}$$

These equations allow to expand the existing vector space with new documents. In this way, new query documents from an unknown author can be transformed into a pseudo-document of the semantic space:

$$query = \boldsymbol{q}^\top \boldsymbol{U}\boldsymbol{\Sigma}^{-1} \tag{6}$$

where $\boldsymbol{q}^\top$ is the term-weighted query. After the query is transformed into the new space, similarity check of documents and/or terms will be possible.

3.4.2 Principal Component Analysis: Mathematically, PCA is adequate if the term frequency distribution in an author profile is Gaussian, linear, and stationary. In this study we have applied PCA to find a suitable representation of the original author profiles, while denoising it by keeping as much information as possible. For that reason, we tested subspace dimensionalities to retain 99.95% and 99.99% variability in the data, but eventually used the entire set of eigenvectors due to insignificant improvement after the retaining. Thus, the dimensionality is reduced to $min(n, m)$, where n is the number of distinct author profiles, and m is the dictionary size.

3.5 *Extreme Learning Machine*

Extreme learning machines (ELM), proposed by Huang *et al.* [63], are feed-forward neural networks with a single hidden layer for classification or regression. They are very fast to learn, compared to other conventional learning methods. In ELM, the initial weights before the hidden layer are randomly assigned and the weights after the hidden layer non-linearity are analytically solved [64]. ELMs were used in text information retrieval before [65], but their usage for author attribution is novel.

In our pipeline for ELM, we optimise the type of the kernel (tested kernels are sigmoid, Gaussian, tanh and multiquadratic), the number of nodes in the hidden layer (λ) (tested in the range of $[100 - 800]$ with step size 10), as well as the mixing coefficient (α) and the width coefficient (ω) (tested for $[0 - 1]$ range with step size of 0.1) on the validation set.

4 Corpora Used in Experimental Evaluation

We have used two Turkish datasets for our evaluations, which we detail in this section. Additionally, we test the generalisation of the

Table 5 *Per user statistics of the two Turkish datasets.*

Corpus	COPA (403 users)					Ekşisözlük (252 users)				
Characteristics	mean	median	minimum	maximum	st. deviation	mean	median	minimum	maximum	st.deviation
# of chat sessions	461	344	200	2,667	349	N/A	N/A	N/A	N/A	N/A
# of entries	5,778	3,111	547	55,534	7,307	491	415	251	2,525	245
# of unique words	3,749	2,350	150	29,396	4,011	9,273	8,517	1,899	35,822	3,266
# of words	19,372	10,221	1,551	190,082	25,224	23,223	20,049	2,892	161,119	13,829
# of letters	70,030	36,430	4,795	718,020	93,678	142,927	121,813	17,876	1,024,682	86,967
# of capital letters	368	0	0	123,278	6194	0.15	0	0	10	0.90
# of emoticons	480	77	0	6,918	1,016	8	3	0	110	14
# of digits	299	125	0	3,263	451	1,271	958	78	9,911	1,150
# of punctuations	1,254	311	0	38,303	2,990	5,890	4,956	620	33,036	3,450

Table 6 *Top-10 words in COPA and Ekşisözlük Databases.*

COPA Database		
Word Used	**Meaning in English**	**Frequency**
slm	abbreviation of "hello"	0.01299
ben	I	0.01159
ne	what	0.01021
sen	you	0.00857
yok	not	0.00835
bu	this	0.00799
ya	or	0.00729
tşk	abbreviation of "thanks"	0.00649
evet	yes	0.00643
tbr	abbreviation of "congrats"	0.00639
Ekşisözlük Database		
Word Used	**Meaning in English**	**Frequency**
bir	one/some	0.02172
de/da	so/also/too/either	0.01627
ve	and	0.01149
bu	this	0.01071
bkz	redirection abbreviation	0.00476
o	he/she/it/that	0.00469
çok	much/many/very	0.00450
için	for/so	0.00434
ne	what	0.00412
ama	but	0.00403

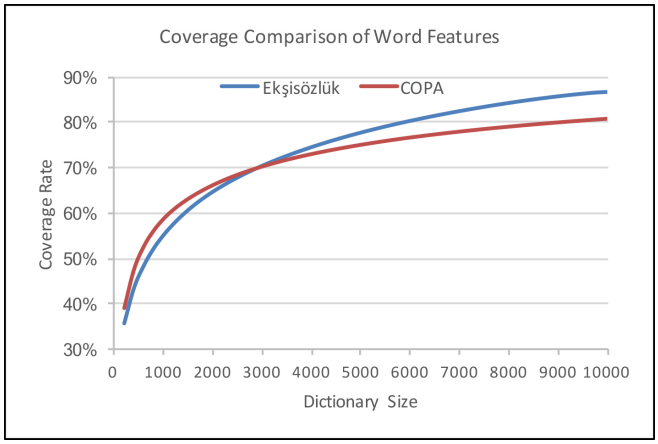


Fig. 3 *Comparison of dictionary coverage: ratio of total number of terms represented by the dictionary subspace to the actual total number of terms in the textual dataset.*

proposed approaches on corpora of English and Portuguese news articles.

4.1 Turkish Corpora

4.1.1 The COPA Database: The first database we have used is the proprietary COPA Database, which consists of demographics, statistics, game records, interactions and complaints of thousands of game players [1]. The database is acquired from a commercial online *Okey* game, over a six months period, and incorporates roughly 100,000 unique players, who played the game at least once. All the player identification information is deleted to protect player privacy.

The database is particular in that messages are always written in a multi-participant fashion (there are always four players in a game of Okey); they are unedited (except for a black-list that contains the most frequently attempted insults); and they are spontaneously produced. The number of chat and game records per player vary greatly. Consequently, we have pre-selected a subset of the dataset for the problem of chat biometrics before any research or modelling took place. We sorted chat participants according to the number of unique words used by each, and eliminated participants who had vocabulary sizes less than 100 unique words. This is a very coarse pre-processing, but people with very limited vocabularies may be easier to identify, and this might positively bias the results. The remaining users are sorted in decreasing order according to the number of active chat sessions, and the most active users are selected for building a chat biometrics benchmark database. We have selected a sub-corpus containing 403 unique users, which is one order of magnitude larger than the most relevant chat biometrics works from the literature. There are several studies that investigated recognition of up to 1,000 authors from their Twitter micro-texts [21, 47], but such texts create different experimental conditions.

4.1.2 The Ekşisözlük Database: Ekşisözlük is a collaborative hypertext "dictionary" in Turkish, based on user-generated content. It is not a dictionary in the strict sense, but contains witty entries on different topics written by over 400,000 registered users.

For this study, we have randomly selected 252 authors to create a test corpus, each of whom having approximately 500 entries on different topics. The entries range from a few characters to hundreds of words.

4.1.3 The Comparison of Turkish Datasets: Per user statistics are given in Table 5: Ekşisözlük users prefer longer words, and use a richer vocabulary, which means more diversity in content. On the other hand, COPA authors use less unique words and have shorter length per word, meaning that they commonly prefer abbreviations.

If we look at the most frequent words in the COPA database, many conversations are specific to games played by the participants and mainly contain greeting and gratitude expressions, as seen in the Table 6. On the other hand in the Ekşisözlük database, pronouns, adjectives, adverbs and conjunction words are mostly preferred by users.

Conversations in the COPA database (in which game participants have many idiosyncratic behaviours including spelling mistakes and shortenings) are dominated by a limited number of unique words with very high frequency. When we look at dictionary coverage (Figure 3), the coverage rate of COPA is higher than that of Ekşisözlük until the number of unique terms (k) reaches 3,000, and it becomes flatter after that. Conversely, the Ekşisözlük database contains relatively more daily life topics and less grammatical errors, and this is reflected in the slope.

4.2 Non-Turkish Corpora

4.2.1 C10 Database: The C10 database contains a subset of English News from the Reuters Corpus Volume I (RCV1), which has over 800,000 manually categorised news. The subset for author attribution is composed of 10 candidate authors, each of whom has 100 texts labelled in Corporate/Industrial (CCAT) group of RCV1 [66]. It was previously used for comparative evaluation of 15 influential author identification methods [67]. For that reason, we report the accuracy of the approaches we propose on the C10 benchmark.

4.2.2 The Portuguese News Database: This dataset is composed of documents extracted from online newspapers and blogs, all written in Brazilian Portuguese [37, 68]. There are 100 authors, each of whom having 30 documents. A total of 3,000 documents are divided into 10 categories, according to 10 different subjects, in a balanced way. The average size of of documents is 2,989 ($\pm 1,531$) bytes. Each document has, on average, 486 tokens and 286 hapax legomena (words occurring just once).

5 Experiments and Results

5.1 Experimental Protocol

We have used 5-fold cross validation in all the experiments, where the text produced by each author is divided into non-overlapping folds. The feature extraction is performed separately for each fold in order to guarantee that the test data is held out of the entire process.

In C10 and Portuguese News datasets, essays are divided into folds for each author without concatenation, so that the results are comparable to other benchmark studies. In other words, each document is an instance for an author in these datasets. On the other hand in COPA, the accumulated chat records of a user in one session are used as an instance. For Ekşisözlük, each individual entry of a user is assumed as an instance. Thus, the instances of a user are concatenated in order to create a more representative profile for that user. While testing, the number of concatenated instances per author was increased from 1 to 50 for COPA and Ekşisözlük in order to analyse how the performance changes with respect to the number of documents per author.

5.1.1 Turkish Normalisation: A small part of the raw data on COPA was normalised by using the web API of a Turkish NLP tool [69], whereby intentional or accidental misspellings were replaced with correct forms. Since Turkish has flexible sentence structure, as well as agglutinative word forms, the normalisation affects the identification performance significantly. For instance the raw sentence "*büttttttüüüüünnnn insnlar e$it dogaaarr*" ("all people are born equal") is normalised as "*bütün insanlar eşit doğar*". Normalisation changes the distribution of the features. Hence, COPA-NORM dataset is created to understand the effect of normalisation on authorship analysis by using a 83 author subset (due to query limitation of the NLP tool).

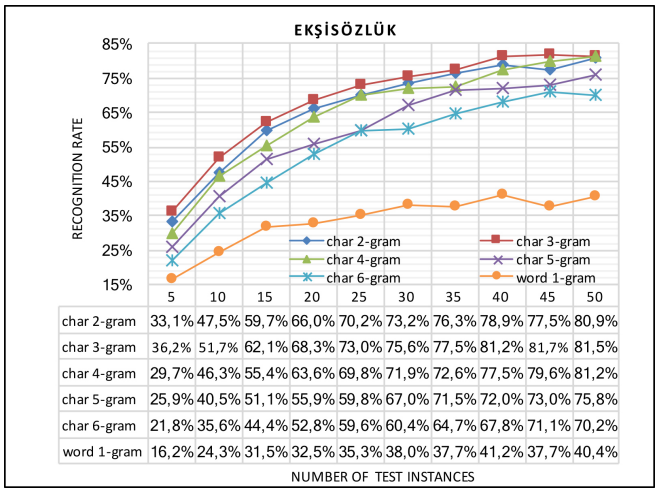
5.2 Fine-tuning on the Pipeline

5.2.1 Effect of Feature Type: The application scenario we mainly focus on is closed-set recognition. Firstly, we have compared author recognition rates of the two Turkish datasets by using character n-gram and word frequency features on the proposed instance-based pipeline. For both databases, 3-gram and 4-gram character features give better recognition rates than their alternatives, as shown in Figure 4. Higher order n-grams for words are not feasible, as they are computationally very expensive.

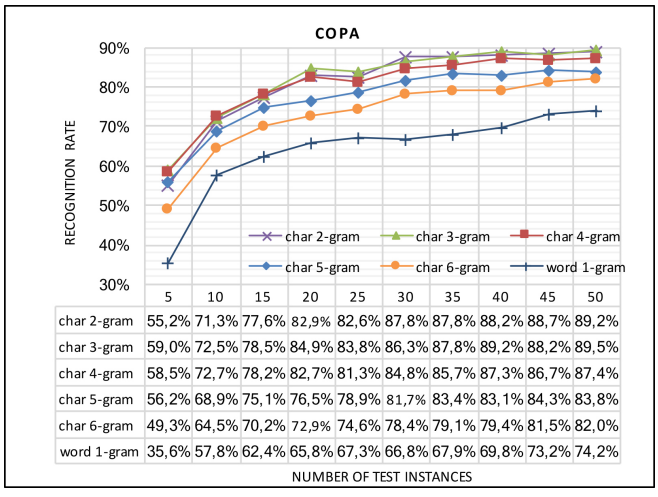
We note that the recognition rates on Ekşisözlük dataset with word frequency features is much lower than with character n-grams, while word frequency of COPA gives similar patterns with character n-grams. The reason is the amount of grammatical mistakes prevalent in social chat (i.e. COPA), which makes misspellings into distinctive indicators for their authors.

5.2.2 Dictionary Size Reduction: If computational complexity is deemed to be important for the system, the dimensionality of the dictionary should be reduced. In this case, how the subspace of the full dictionary is determined is an optimisation issue. We illustrate this on the C10 dataset, where global frequent ranking (GFR), local frequent ranking (LFR), and local distinctive ranking (LDR) are compared with each other on the proposed instance-based approach, while the number of unique terms (k) is increased from 1,000 to 10,000. As seen in Figure 5, LFR and LDR are more robust compared to GFR (the lines above the bars) on changing dictionary coverage (the blue bars). Dual comparisons among them with respect to 1-tailed t-test shows that LFR and LDR are significantly better than GFR with p values of 0.018 and 0.001 respectively. On the other hand, the results show that when dictionary size reaches 5,000 (39% of all terms existing in the dataset), ranking methods for representing VSM have no superiority to each other (in terms of paired t-tests, $p > 0.05$).

5.2.3 Feature Weighting: Weighting schemes used in the experiments are **i)** term frequency - inverse document frequency (TF-IDF), **ii)** sublinear term frequency - inverse document frequency (sTF-IDF), and **iii)** entropy - logarithmic term frequency (Entropy-Log). For author attribution, Layton *et al*. used TF-IDF weighting in the inverse author frequency (IAF) scheme and reached promising results [54]. In a similar manner, VSM weighting with sTF-IDF gives better cross-validation accuracy results on C10 database, compared to the test pipeline without weighting. On the other hand,



	5	10	15	20	25	30	35	40	45	50
char 2-gram	33,1%	47,5%	59,7%	66,0%	70,2%	73,2%	76,3%	78,9%	77,5%	80,9%
char 3-gram	36,2%	51,7%	62,1%	68,3%	73,0%	75,6%	77,5%	81,2%	81,7%	81,5%
char 4-gram	29,7%	46,3%	55,4%	63,6%	69,8%	71,9%	72,6%	77,5%	79,6%	81,2%
char 5-gram	25,9%	40,5%	51,1%	55,9%	59,8%	67,0%	71,5%	72,0%	73,0%	75,8%
char 6-gram	21,8%	35,6%	44,4%	52,8%	59,6%	60,4%	64,7%	67,8%	71,1%	70,2%
word 1-gram	16,2%	24,3%	31,5%	32,5%	35,3%	38,0%	37,7%	41,2%	37,7%	40,4%



	5	10	15	20	25	30	35	40	45	50
char 2-gram	55,2%	71,3%	77,6%	82,9%	82,6%	87,8%	87,8%	88,2%	88,7%	89,2%
char 3-gram	59,0%	72,5%	78,5%	84,9%	83,8%	86,3%	87,8%	89,2%	88,2%	89,5%
char 4-gram	58,5%	72,7%	78,2%	82,7%	81,3%	84,8%	85,7%	87,3%	86,7%	87,4%
char 5-gram	56,2%	68,9%	75,1%	76,5%	78,9%	81,7%	83,4%	83,1%	84,3%	83,8%
char 6-gram	49,3%	64,5%	70,2%	72,9%	74,6%	78,4%	79,1%	79,4%	81,5%	82,0%
word 1-gram	35,6%	57,8%	62,4%	65,8%	67,3%	66,8%	67,9%	69,8%	73,2%	74,2%

Fig. 4 *Comparison of recognition rate and lexical features for Ekşisözlük and COPA. (Dictionary size: 5,000 with local frequent ranking, no weighting, ELM params: $\lambda = 250$, $\alpha = 0,5$, $\omega = 0,5$.)*

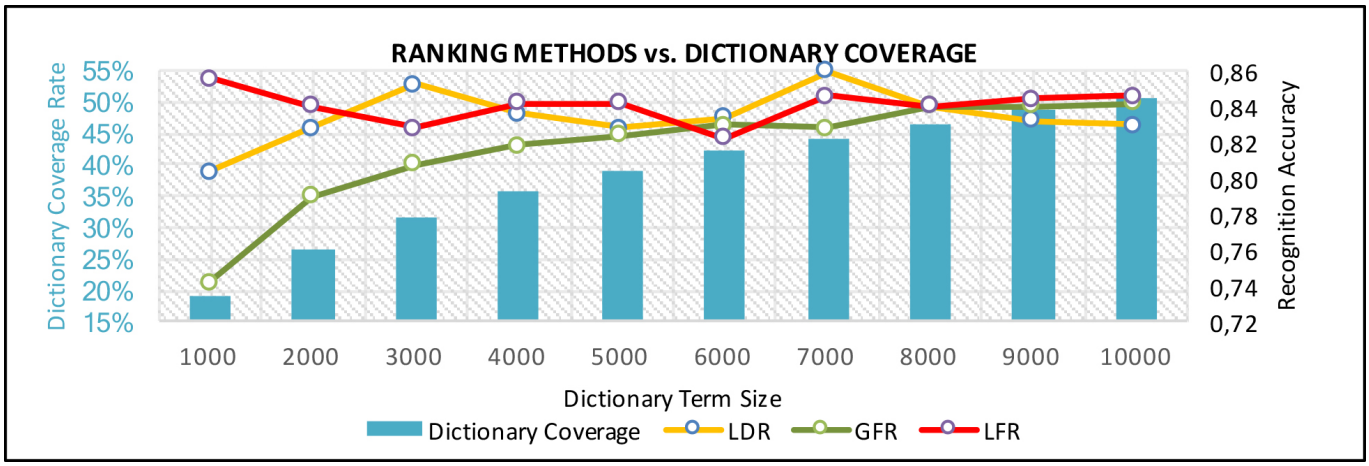


Fig. 5 *Average cross-validation accuracies with different rankings & dictionary size for C10 dataset (6-gram char. features, TF-IDF weighting, ELM parameters:* $\lambda = 230$, $\alpha = 0,7$, $\omega = 0,9$*).*

TF-IDF outperforms other weighting methods for Ekşisözlük dataset which means that the performance of VSM weighting is strongly dependent upon dataset, as shown in Table 7. Moreover, we observe on the COPA dataset that the test pipeline with weighting methods do not outperform the non-weighted approach in case of data concatenation for each author. The reason for this is that global weighting reduces the effect of terms used by many authors, and as corpus size is increased, even rare words are used by multiple authors, thus reducing their discriminativeness. For instance, about 85% of terms in COPA are weighted with 0 in our experiments. On the other hand, if a training corpus has a limited amount of text for each author (as in the case of C10 and Ekşisözlük), the weighting scheme may lead to remarkable improvements on the recognition rate.

5.2.4 Supervised Classification: We used ELM in our instance-based pipeline, mainly due to its training speed. In this section we compare it with Naive Bayes, Multi-layer Perceptron (MLP), and Support Vector Machine classifiers, which are commonly used in the literature. In order to observe performance variations on language domain changes, we conduct experiments on both C10 and Ekşisözlük datasets.

During grid search for parameter optimisation, character n-gram type ($n = 2, 3, 4, 5, 6$), weighting methods (TF-IDF, sTF-IDF, Log-Entropy, and no weighting), existence of LSA (yes/no), number of unique terms (k) in dictionary (in between 1,000-10,000 with step size of 1,000), and hyper-parameters of each classifier are optimised to get the best performance.

For ELM, optimisation is made as mentioned in Section 3.5 by using the implementation of Lambert'13 [70] on Python 2.7.3. For the rest of the classifiers, Python's scikit-learn library is used with the optimisation of following parameters: i) for SVM, kernel type (RBF or linear) and the penalty parameter ($C = [0 - 1]$ with step size of 0.1), ii) for Naive Bayes, model type (multivariate Bernoulli, Gaussian, or multinomial) and smoothing parameter ($\alpha = [0 - 1]$ with step size of 0.1), and iii) for MLP, activation function (rectified linear unit, logistic, tanh) and the number of hidden units ($\lambda = [100 - 800]$ with the step size of 10). As seen in Table 8, ELM gives a higher average accuracy than other supervised learning methods on the Ekşisözlük dataset. On the other hand, both MLP and ELM have similar accuracy, outperforming other learning methods on the C10 dataset.

5.3 Benchmarking on the Literature

We validate our recognition methodology both on Turkish and non-Turkish authorship problems. Table 9 compares our proposed approaches with several works from the literature [17, 26, 31, 54] on the four datasets described before. We report accuracy and macro-average F_1-Score to account for imbalanced labels.

For Turkish datasets, COPA and Ekşisözlük, where author entries are concatenated to create longer texts, our profile-based methodology outperforms both profile-based and instance based approaches in the literature, including previous efforts in Turkish. On the other hand, for non-Turkish datasets, where each essay is treated as an instance, very promising results are obtained with our instance-based approach. On the C10 dataset, the proposed methodology improves on the 15 author identification methods reproduced in the works of Potthast *et al.*, where the best recognition accuracy was noted as 76.6% [67]. We obtain $87.6\%(\pm 2.3\%)$ cross validation accuracy and 81.2% test set accuracy on the C10 dataset. Similar strong results were obtained on the Portuguese News database.

We compare the performances of the tested approaches, and report pairwise significance in Table 10. On the Ekşisözlük dataset, PBA was observed to perform significantly better than the other authorship attribution approaches tested, and IBA is slightly better than these. Not surprisingly, RLP gives the worst result on Ekşisözlük, even though it has a good performance on COPA. The reason for this is the particular writing style imposed by Ekşisözlük, which suppresses the authors from having totally divergent styles. Since RLP is based on dissimilarity measurement of local distinctive features, it does not extract sufficiently diverse features for author profiles.

While the number of concatenated documents is increased in a query from 1 to 50, these methods were observed to reach a saturation level after 25 instances, as shown in Figure 6. The Naive Bayes classifier was slower to reach these levels of accuracy. Nevertheless, if only one instance is queried, the proposed PBA and IBA methods are significantly superior to other methods: they have 34.9% and 31.5% Rank-1 accuracy, respectively, while these rates are only 29.2%, 28.4% and 25.6% for SCAP, CNG and RLP methods, respectively. Similar to the rest of the experiments in the study, the choice of n-gram ($n = 2, 3, 4, 5, 6$) and the dictionary size ($k = [1,000 - 10,000]$ with step size of $1,000$) are optimised separately for each baseline method (CNG, RLP, and SCAP, respectively).

Results shown in Figure 6 illustrate that the proposed PBA and IBA are more robust to domain changes on authorship attribution, while the accuracy of RLP and CNG are not stable to such changes.

5.4 Limited Text for Recognition

We have investigated the performance of the proposed PBA under the condition that very limited text exists per gallery author. Figure 7 illustrates the cumulative match characteristic (CMC) curves for the case where a single document is used per author. According to these results, the performance of SCAP is similar to PBA under very limited text conditions, and RLP performs worst. Very similar patterns are observed for both datasets.

Table 7 *Comparison of weighting schemes with changing character n-grams on C10 and Ekşisözlük datasets. (Dictionary size: 5,000 with local frequent ranking, ELM parameters: $\lambda = 230$, $\alpha = 0,7$, $\omega = 0,9$)*

Weighting Schemes	C10 Database				Ekşisözlük Database			
	n=3	n=4	n=5	n=6	n=3	n=4	n=5	n=6
No Weighting	0.834±0.027	0.848±0.034	0.838±0.031	0.838±0.021	0.646±0.060	0.668±0.054	0.644±0.062	0.605±0.062
TF-IDF	0.820±0.026	0.844±0.029	0.846±0.031	0.844±0.031	**0.719**±0.066	**0.715**±0.070	**0.683**±0.069	**0.654**±0.071
sTF-IDF	**0.856**±0.029	**0.850**±0.036	**0.864**±0.027	**0.858**±0.032	0.581±0.061	0.546±0.070	0.515±0.061	0.499±0.056.
Log-Entropy	0.826±0.037	0.840 ±0.038	0.850±0.036	0.848±0.036	0.578±0.058	0.516±0.062	0.517±0.060	0.509±0.064

Table 8 *Cross validation accuracy for classifiers, and the parameters that give the best recognition results.*

	Ekşisözlük Database		C10 Database	
	Best Accuracy	**Optimum Parameters**	**Best Accuracy**	**Optimum Parameters**
ELM	0.**864**±0.027	ELM:Multiquadric, $\lambda = 240$, $\alpha = 0,2$, $\omega = 1,0$ $n = 4$, $k = 3,000$ weight:TF-IDF, LSA:Yes	**0.876**±0.023	ELM:Multiquadric, $\lambda = 250$, $\alpha = 0,5$, $\omega = 0,7$ $n = 6$, $k = 5,000$, weight:sTF-IDF, LSA:Yes
SVM	0.700±0.030	SVM: linear model, $C = 1$, $n = 5$ $k = 2,000$ weight:TF-IDF, LSA:Yes	0.858±0.029	SVM: linear model, $C = 0.9$, $n = 5$ $k = 5,000$ weight:sTF-IDF, LSA:Yes
NB	0.732±0.055	NB: multivariate Bernoulli, $\alpha = 1$, $n = 5$ $k = 2,000$, weight:TF-IDF, LSA:Yes	0.844±0.022	NB: multivariate Bernoulli, $\alpha = 1$, $n = 5$ $k = 3,000$, weight:No, LSA:No
MLP	0.852±0.038	MLP: Rectified linear unit, $\lambda = 210$, $n = 4$ $k = 4,000$, weight:TF-IDF, LSA:No	**0.876**±0.035	MLP: Rectified linear unit, $\lambda = 240$, $n = 5$ $k = 5,000$, weight:sTF-IDF, LSA:No

Table 9 *Comparison of the proposed approaches and leading approaches from the literature.*

	Instance-based Approaches				Profile-Based Approaches							
	Proposed IBA		NB Classifier [17]		Proposed PBA		SCAP [31]		CNG [26]		RLP [54]	
	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy
COPA	94.9%	97.2%	89.8%	94.7%	**99.2%**	**98.5%**	72.2%	76.8%	91.5%	96.1%	94.6%	95.8%
Ekşisözlük	83.1%	86.0%	81.2%	82.7%	**85.0%**	**87.9%**	72.7%	72.8%	71.9%	71.5%	49.7%	51.6%
C10	**81.8%**	**81.2%**	76.9%	77.2%	71.3%	73.2%	73.4%	74.2%	71.4%	72.2%	69.6%	70.6%
Portuguese News	**83.7%**	**83.3%**	79.6%	75.9%	82.5%	81.9%	75.9%	73.3%	75.4%	74.5%	73.9%	72.3%

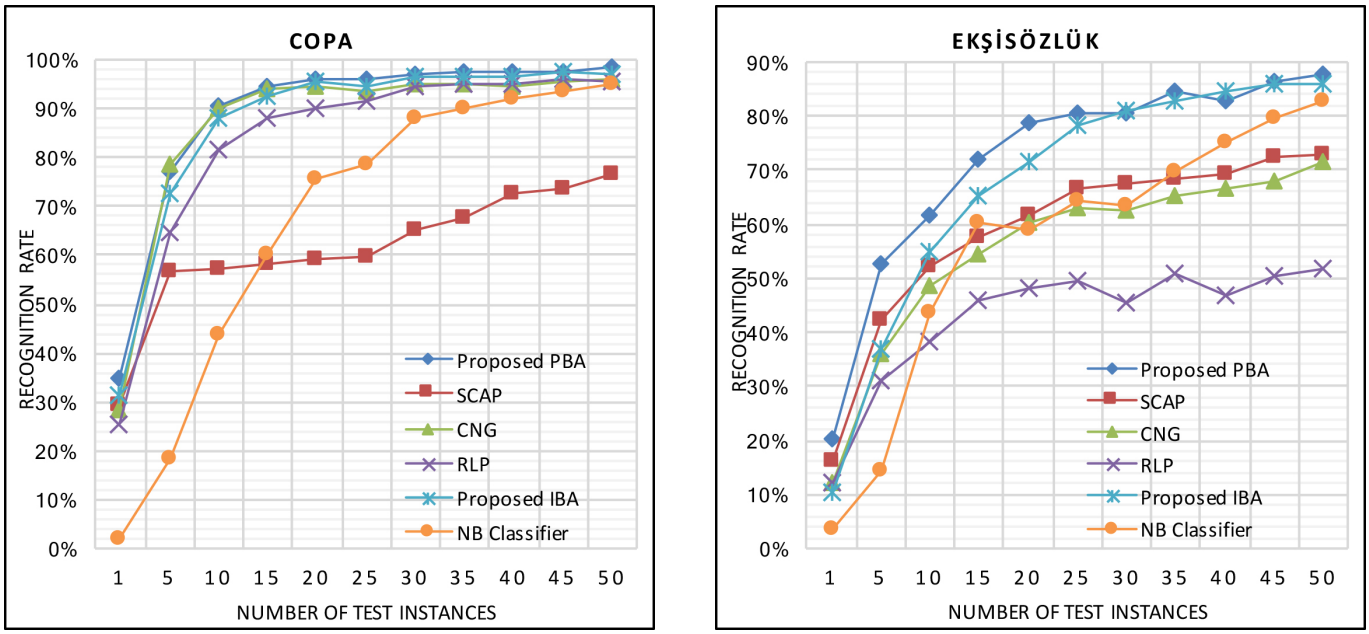


Fig. 6. ***Recognition rate vs. number of query instances on Ekşisözlük and COPA.***

Table 10 *Dual comparison of all approaches with respect to 1-tailed t-test (Significant p-values which are less than 0.05 are shown bold)*

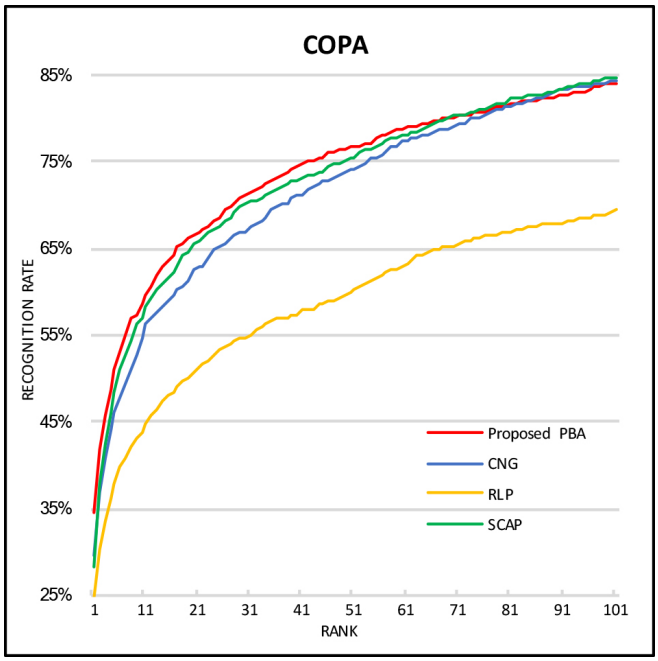
	COPA Dataset					Ekşisözlük Dataset				
	P-IBA	SCAP	CNG	RLP	NB	P-IBA	SCAP	CNG	RLP	NB
Proposed PBA	0.421	**7.3e-4**	0.379	0.277	**0.027**	0.389	**0.050**	**0.024**	**6.9e-4**	**0.049**
Proposed IBA		**0.001**	0.459	0.347	**0.038**		0.116	0.060	**0.002**	0.092
SCAP			0.998	0.995	0.633			0.241	**0.007**	0.267
CNG				0.381	**0.041**				**0.044**	0.419
RLP					0.068					0.870

5.5 Text Normalisation

The last test for Turkish chat records is conducted on the COPA-NORM dataset, by using character 4-gram features. Raw and Turkish normalised versions of the corpus are compared on profile-based and instance-based authorship attributions proposed in this study, as shown in Figure 8. Normalisation leads to accuracy loss as expected, since misspellings or typos are some of the most important features for identification [4]. On the other hand, the accuracy loss due to normalisation becomes insignificant with the increase of chat instances.

6 Conclusions

In this paper, we have proposed an approach for authorship recognition within the context of chat biometrics. We performed tests with a



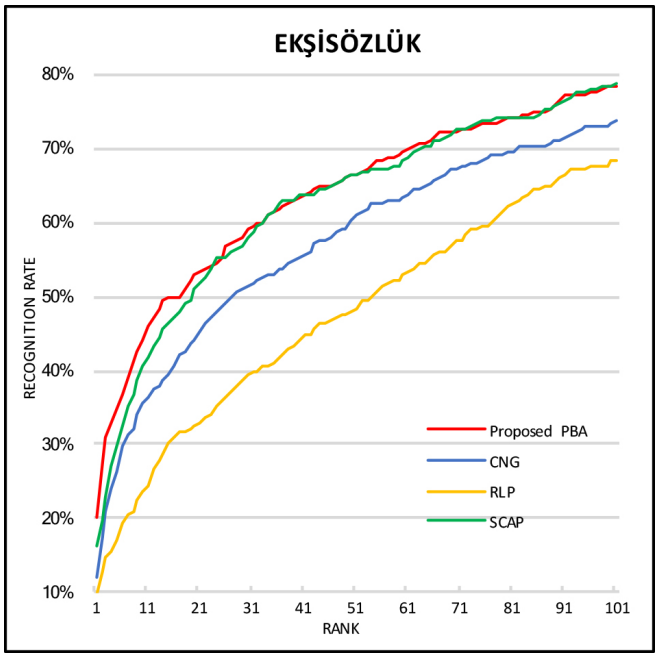
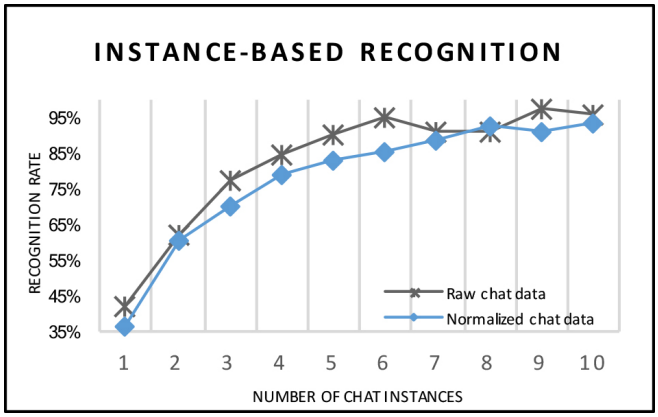


Fig. 7. ***CMC curves for COPA (left) and Ekşisözlük (right) under the assumption of only one text for each gallery author.***



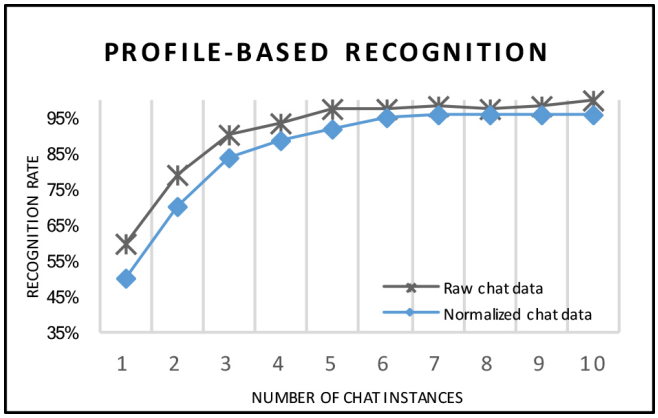


Fig. 8. ***Comparison of recognition rates for COPA-NORM before and after the normalisation on the proposed approaches.***

large database of multiparty chat records in Turkish, which is available upon request for academic purposes, and with a novel database collected from the largest Turkish online social network. We further validated our proposed approach on news datasets in Portuguese and English.

Our results illustrate that domain-specific optimisation of dictionary size via local ranking of terms, and LSA/PCA projection on the feature set are both important for obtaining accurate systems. We contrasted word and character based features, as well as effects of feature weighting schemes. Character based features appear to be more scalable for this problem, and produced better results. We should remark here that LSA/PCA projections could mask stylistic features, such as word choices, by for instance grouping synonyms into a single topic. Such stylistic features could potentially be relevant for identifying authors of literary texts. Indeed, [71] has introduced an approach to analyse style through synonyms, by picking words with many synonyms, and by checking the author's choices for these words. However, the chat domain is characterised by a limited vocabulary, as well as limited amount of text, where semantic analysis may not be employed [21]. Furthermore, correct detection of stylistic elements requires text normalisation, which we have shown to impoverish detection results. Consequently, we have not investigated the use of such stylistic features in detail.

We tested the robustness of the approach to domain variations, by means of the C10 and Portuguese News datasets. We have reached rank-1 recognition rates up to 98.5% and 87.9% on COPA (403 classes) and Ekşisözlük (252 classes) datasets with the profile-based approach. On the other hand, 81.2% and 83.3% accuracy rates are reached on Portuguese (100 classes) and English (10 classes) news datasets with the instance-based approach. These results imply that profile-based approach is better for author attribution on informal chat datasets, while instance-based author attribution method outperforms on well-structured and formal textual data. Our results indicate that for moderately sized closed sets (i.e. up to 1000 authors), and with a fairly small amount of query text, it is possible to identify authors from their online chat communications.

7 Acknowledgements

This work is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with project number 114E481 and EU COST IC1106. We thank Gülşen Eryiğit for allowing us the use of the ITU NLP tools.

8 References

1 K. Balci, A. A. Salah. "Automatic Analysis and Identification of Verbal Aggression and Abusive Behaviors for Online Social Games", *Computers in Human Behavior*, **53**, pp. 517 – 526, (2015).

2 K. Balci, A. A. Salah. "Automatic Classification of Player Complaints in Social Games", *IEEE Trans. on Computational Intelligence and AI in Games*, **9(1)**, pp. 103–108, (2017).
3 T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, F. Can. "Chat Mining: Predicting User and Message Attributes in Computer-Mediated Communication", *Information Processing & Management*, **44(4)**, pp. 1448–1466, (2008).
4 R. S. Kuzu, K. Balci, A. A. Salah. "Authorship Recognition in a Multiparty Chat Scenario", *4th IEEE Int. Conf. on Biometrics and Forensics*, (2016).
5 A. Gray, P. Sallis, S. Macdonell. "Software Forensics: Extending Authorship Analysis Techniques to Computer Programs", *Proc. IAFL*, (1997).
6 T. C. Mendenhall. "The Characteristic Curves of Composition", *Science*, pp. 237–249, (1887).
7 F. Mosteller, D. L. Wallace. "Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers", *Journal of the American Statistical Association*, **58(302)**, pp. 275–309, (1963).
8 E. Stamatatos. "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, **60(3)**, pp. 538–556, (2009).
9 R. Zheng, J. Li, H. Chen, Z. Huang. "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques", *Journal of the American Society for Information Science and Technology*, **57(3)**, pp. 378–393, (2006).
10 A. K. Jain, A. Ross, S. Prabhakar. "An Introduction to Biometric Recognition", *IEEE Trans. on Circuits and Systems for Video Technology*, **14(1)**, pp. 4–20, (2004).
11 E. Stamatatos, N. Fakotakis, G. Kokkinakis. "Automatic Text Categorization in Terms of Genre and Author", *Computational Linguistics*, **26(4)**, pp. 471–495, (2000).
12 S. M. Zu Eissen, B. Stein, M. Kulig. "Plagiarism Detection Without Reference Collections", *Advances in Data Analysis*, pp. 359–366, (Springer, 2007).
13 P. Juola. "Authorship Attribution for Electronic Documents", *Advances in Digital Forensics II*, pp. 119–130, (2006).
14 O. De Vel, A. Anderson, M. Corney, G. Mohay. "Mining E-mail Content for Author Identification Forensics", *ACM Sigmod Record*, **30(4)**, pp. 55–64, (2001).
15 C. Sanderson, S. Guenter. "On Authorship Attribution via Markov Chains and Sequence Kernels", *Proc. ICPR*, volume 3, pp. 437–440, (2006).
16 T. Tufan, A. K. Görür. "Author Identification for Turkish Texts", *Cankaya University Journal of Arts and Sciences*, **1(7)**, (2007).
17 M. F. Amasyalı, B. Diri. "Automatic Turkish Text Categorization in Terms of Author, Genre and Gender", *Natural Language Processing and Information Systems*, pp. 221–226, (2006).
18 B. Diri, M. Amasyalı. "Automatic Author Detection for Turkish Texts", *Proc. ICANN/ICONIP*, pp. 138–141, (2003).
19 L. Šarkute, A. Utka. "The Effect of Author Set Size in Authorship Attribution for Lithuanian", *Proc. NODALIDA*, p. 87, (2015).
20 N. Potha, E. Stamatatos. "A Profile-based Method for Authorship Verification", *Artificial Intelligence: Methods and Applications*, pp. 313–326, (Springer, 2014).
21 A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. Carvalho, E. Stamatatos. "Authorship attribution for social media forensics", *IEEE Trans. on Information Forensics and Security*, **12(1)**, pp. 5–33, (2017).
22 S. Ruder, P. Ghaffari, J. G. Breslin. "Character-level and multi-channel convolutional neural networks for large-scale authorship attribution", *arXiv preprint arXiv:1609.06686*, (2016).
23 S. Wang, E. Ferracane, R. J. Mooney. "Leveraging discourse information effectively for authorship attribution", *arXiv preprint arXiv:1709.02271*, (2017).
24 Y. Sari, A. Vlachos, M. Stevenson. "Continuous n-gram representations for authorship attribution", *Proc. EACL*, pp. 267–273, (2017).
25 P. Shrestha, S. Sierra, F. A. González, P. Rosso, M. Montes-y Gómez, T. Solorio. "Convolutional neural networks for authorship attribution of short texts", *Proc. EACL*, pp. 669–674, (2017).
26 V. Kešelj, F. Peng, N. Cercone, C. Thomas. "N-Gram-Based Author Profiles for Authorship Attribution", *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING*, volume 3, pp. 255–264, (2003).
27 P. Clough. "Old and new challenges in automatic plagiarism detection", *National Plagiarism Advisory Service*, pp. 391–407, (2003).
28 Y. Zhao, J. Zobel, P. Vines. "Using Relative Entropy for Authorship Attribution", *Information Retrieval Technology*, pp. 92–105, (Springer, 2006).
29 C. Sanderson, S. Guenter. "Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation", *Proc. EMNLP*, pp. 482–491, (2006).
30 P. M. McCarthy, G. A. Lewis, D. F. Dufty, D. S. McNamara. "Analyzing Writing Styles with Coh-Metrix", *Florida AI Research Society Conf.*, pp. 764–769, (2006).
31 G. Frantzeskou, E. Stamatatos, S. Gritzalis, C. E. Chaski, B. S. Howald. "Identifying Authorship by Byte-Level N-Grams: The Source Code Author Profile (SCAP) Method", *International Journal of Digital Evidence*, **6(1)**, pp. 1–18, (2007).
32 D. Estival, T. Gaustad, S. B. Pham, W. Radford, B. Hutchinson. "Author Profiling for English Emails", *Proc. PACLING*, pp. 263–272, (2007).
33 S. Argamon, M. Koppel, J. W. Pennebaker, J. Schler. "Automatically Profiling the Author of an Anonymous Text", *Comm. ACM*, **52(2)**, pp. 119–123, (2009).
34 M. Koppel, J. Schler, S. Argamon. "Authorship Attribution in The Wild", *Language Resources and Evaluation*, **45(1)**, pp. 83–94, (2011).
35 T. Solorio, S. Pillay, S. Raghavan, M. Montes-y Gómez. "Modality Specific Meta Features for Authorship Attribution in Web Forum Posts", *IJCNLP*, pp. 156–164, (2011).
36 H. J. Escalante, T. Solorio, M. Montes-y Gómez. "Local Histograms of Character N-grams for Authorship Attribution", *Proc. ACL*, pp. 288–298, (2011).
37 W. Oliveira Jr, E. Justino, L. Oliveira. "Authorship Attribution of Documents Using Data Compression as a Classifier", *Proc. World Congress on Engineering and Computer Science*, volume 1, (2012).
38 R. Layton, P. Watters, R. Dazeley. "Recentred Local Profiles for Authorship Attribution", *Natural Language Engineering*, **18(03)**, pp. 293–312, (2012).
39 J. Savoy. "Authorship Attribution Based on Specific Vocabulary", *ACM Trans. on Information Systems (TOIS)*, **30(2)**, p. 12, (2012).
40 M. Cristani, G. Roffo, C. Segalin, L. Bazzani, A. Vinciarelli, V. Murino. "Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging", *Proc. ACM Multimedia*, pp. 1121–1124, (2012).
41 S. Seidman. "Authorship Verification Using the Impostors Method", *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*, (2013).
42 G. Inches, M. Harvey, F. Crestani. "Finding Participants in a Chat: Authorship Attribution for Conversational Documents", *Social Computing (SocialCom), 2013 International Conference on*, pp. 272–279, (IEEE, 2013).
43 J. V. Monaco, J. C. Stewart, S.-H. Cha, C. C. Tappert. "Behavioral Biometric Verification of Student Identity in Online Course Assessment and Authentication of Authors in Literary Works", *Proc. IEEE BTAS*, (2013).
44 G. Roffo, M. Cristani, L. Bazzani, H. Minh, V. Murino. "Trusting Skype: Learning the Way People Chat for Fast User Recognition and Verification", *CVPR Workshops*, pp. 748–754, (2013).
45 M. L. Brocardo, I. Traore, S. Saad, I. Woungang. "Authorship Verification for Short Messages using Stylometry", *Proc. IEEE CITS*, (2013).
46 F. Iqbal, H. Binsalleeh, B. C. Fung, M. Debbabi. "A Unified Data Mining Solution for Authorship Analysis in Anonymous Textual Communications", *Information Sciences*, **231**, pp. 98–112, (2013).
47 R. Schwartz, O. Tsur, A. Rappoport, M. Koppel. "Authorship Attribution of Micro-Messages", *Conference on Empirical Methods in Natural Language Processing*, volume 3, pp. 1880–1891, (2013).
48 G. K. Mikros, K. Perifanos. "Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles.", *AAAI Spring Symposium: Analyzing Microtext*, (2013).
49 T. Qian, B. Liu, L. Chen, Z. Peng. "Tri-Training for Authorship Attribution with Limited Training Data", *Association of Computational Linguistics (2)*, pp. 345–351, (2014).
50 Y. Seroussi, I. Zukerman, F. Bohnert. "Authorship Attribution with Topic Models", *Computational Linguistics*, **40(2)**, pp. 269–310, (2014).
51 S. Segarra, M. Eisen, A. Ribeiro. "Authorship Attribution through Function Word Adjacency Networks", *IEEE Trans. on Signal Processing*, **63(20)**, pp. 5464–5478, (2015).
52 R. Overdorf, R. Greenstadt. "Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution", *Proc. Privacy Enhancing Technologies*, **(3)**, pp. 155–171, (2016).
53 E. Stamatatos. "Authorship attribution using text distortion", *Proc. EACL*, pp. 1138–1149, (2017).
54 R. Layton, S. McCombie, P. Watters. "Authorship Attribution of IRC Messages Using Iinverse Author Frequency", *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*, pp. 7–13, (IEEE, 2012).
55 M. Brennan, S. Afroz, R. Greenstadt. *ACM Trans. on Information and System Security*, **15(3)**, p. 12, (2012).
56 M. Eder. "Does size matter? authorship attribution, small samples, big problem", *Digital Scholarship in the Humanities*, **30(2)**, pp. 167–182, (2015).
57 E. Aydın Oktay, K. Balcı, A. A. Salah. "Automatic assessment of dimensional affective content in Turkish multi-party chat messages", *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, pp. 19–24, (ACM, 2015).
58 G. Salton, M. J. McGill. *Introduction to Modern Information Retrieval*, (McGraw-Hill, Inc., 1986).
59 P. Soucy, G. W. Mineau. "Beyond tfidf weighting for text categorization in the vector space model", *IJCAI*, volume 5, pp. 1130–1135, (2005).
60 R. S. Kuzu, A. Haznedaroğlu, M. L. Arslan. "Topic Identification for Turkish Call Center Records", *Proc. IEEE SIU*, (2012).
61 N. Ali, M. Price, R. Yampolskiy. "BLN-Gram-TF-ITF as a New Feature for Authorship Identification", *Academy of Science and Engineering (ASE) BIG-DATA/SOCIALCOM/CYBERSECURITY Conference*, (2014).
62 T. K. Landauer, P. W. Foltz, D. Laham. "An Introduction to Latent Semantic Analysis", *Discourse Processes*, **25(2-3)**, pp. 259–284, (1998).
63 G.-B. Huang, Q.-Y. Zhu, C.-K. Siew. "Extreme Learning Machine: Theory and Applications", *Neurocomputing*, **70(1)**, pp. 489–501, (2006).
64 G.-B. Huang, H. Zhou, X. Ding, R. Zhang. "Extreme Learning Machine for Regression and Multiclass Classification", *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **42(2)**, pp. 513–529, (2012).
65 W. Zheng, Y. Qian, H. Lu. "Text Categorization Based on Regularization Extreme Learning Machine", *Neural Computing and Applications*, **22(3-4)**, pp. 447–456, (2013).
66 D. D. Lewis, Y. Yang, T. G. Rose, F. Li. "RCV1: A New Benchmark Collection for Text Categorization Research", *Journal of Machine Learning Research*, **5(Apr)**, pp. 361–397, (2004).
67 M. Potthast, S. Braun, T. Buz, F. Duffhauss, F. Friedrich, J. M. Gülzow, J. Köhler, W. Lötzsch, F. Müller, M. E. Müller, et al. "Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval", *European Conference on Information Retrieval*, pp. 393–407, (Springer, 2016).
68 P. J. Varela. "O Uso de Atributos Estilométricos na Identificação da Autoria de Textos", Ph.D. thesis, Pontifícia Universidade Católica do Paraná, (2010).
69 G. Eryigit. "ITU Turkish NLP Web Service", *Proc. EACL*, (2014).
70 D. C. Lambert. ELM v0.3 edition, (2013), [Online]. Available: https://github.com/dclambert/Python-ELM.
71 J. H. Clark, C. J. Hannon. "A classifier system for author recognition using synonym-based features", *Mexican Int. Conf. on AI*, pp. 839–849, (2007).