AUTHORSHIP RECOGNITION IN A MULTIPARTY CHAT SCENARIO

Rıdvan Salih Kuzu, Koray Balcı, Albert Ali Salah[*]

Boğaziçi University, Department of Computer Engineering
34342 Bebek - Istanbul, TURKEY
ridvan.salih@boun.edu.tr, balci@cmpe.boun.edu.tr, salah@boun.edu.tr

ABSTRACT

Users of online social networks often use multiple identities. This paper investigates the possibility of identifying a user from his or her chat behavior in such a setting. We have collected a large corpus of multiparty chat records in Turkish, obtained from a multiplayer game database. The most active 978 users are selected according to their participation in game chat sessions. This corpus is used in a biometric identification experiment where we seek each user among a gallery of users. Character matrices for each player are used as features, and re-centered local profiles and cosine similarity measure are preferred as identification methods. We systematically assess the effect of text normalization on identification. We report comparative results, the best of which reach around 75% rank-1 accuracy for a gallery size of 978.

***Index Terms*—** Chat biometrics; Multiparty chat; Chat mining; Authorship recognition; Machine learning; Text classification; Text information retrieval

1. INTRODUCTION

Computer-mediated communication with text messages has become popular with the increase of internet era. Instant messaging applications on mobile devices such as WhatsApp, Line, Viber, Skype, SnapChat have received widespread attention, and multiplayer games with common chat rooms are popular among young people. These internet based services constantly generate large amount of text data, which can be processed by applications of sentiment analysis and user analytics. The informal nature of these texts, their unordered structure, and the large amount of spelling mistakes bring additional challenges to the typical natural language processing based analysis.

Textual messaging can be in the form of one-to-one communication, or it can involve group messaging, which can also be referred to as multiparticipant chat [1]. Multiparticipant chat has numerous application scenarios including technical support, recreation business, online courses, collaborative learning and gaming. A recently investigated application scenario involves verbal aggression and abuse during chat in social online games [2]. Owners of such online platforms usually moderate social interactions, and verbal abuse and cyberbullying typically result in temporary or permanent banning from the platform. However, creation of multiple accounts is a frequent practice. In this paper we investigate the possibility of identifying a person from his or her chat behavior.

Writing style is unique for everybody, and some identity-related cues remain even if the individual consciously attempts to change the writing style [3]. This issue was investigated in the context of authorship recognition, which seeks to identify the author of a piece of text from 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.

This study uses data acquired from the chat interface of a multiplayer online game. In such games, some users who are blocked by administrators for various reasons (such as cheating, foul language, hate speech, abusive behaviors) 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 containing 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 these malicious users in multiparticipant chat environments by means of extracting relevant features and supervised classification techniques. In our approach, we apply and compare several methods to match users to a gallery by their chat records. In the literature, methods developed for matching personal text content have been mostly evaluated with Indo-European lan-

[*] This work is supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with project number 114E481 and by the Turkish Ministry of Development under the TAM Project number DPT2007K120610. Please cite the paper as Kuzu, R.S., K. Balcı, A.A. Salah, "Authorship Recognition in a Multiparty Chat Scenario," Int. Workshop on Biometrics and Forensics, Limassol, March 2016.

guages. We test our approach with documents that have Turkish chat content, which bring additional challenges due to the agglutinative nature of the language (i.e. many postfixes are applied on word roots). Text analysis in agglutinative languages includes a normalization step to isolate the roots of the words, which we additionally assess in the context of chat biometrics.

The rest of the paper is organized as follows. Section 2 overviews the problem and related work. Section 3 describes the proposed method for identifying a person given his or her chat records, baselines, and the dataset used. Section 4 reports our experimental results on the COPA Database, as well as the results of the approach on a standard authorship identification benchmark as a sanity check. Section 5 concludes the paper.

2. RELATED WORK

2.1. Authorship recognition

Every authorship recognition (or identification) problem contains a training corpus in which there is a set of text samples for candidate authors and a test corpus of 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 [4].

Concatenating training texts per author in one single text file is known as the profile-based approach. 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 approaches, on the other hand, consider 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, concatenation of the text (as in profile-based approaches) may help to create a sufficiently long document for capturing the author's style [5]. Performance in this domain depends on identification methods, as well as pre-processing techniques, document set sizes, language characteristics, and feature sets. In terms of used features, character N-grams, word tokens, term frequency - inverse term frequency (TF-ITF), distribution based similarity features are typically used. We summarize some common identification and attribution approaches in terms of feature extraction and matching methods in Table 1. Features are categorized based on different attributes of text: lexical, syntactic, semantic and character based features, respectively [6]. Some of the most commonly used features are listed in Table 2.

There are a few important studies related to chat biometrics on texts in English. Inches et al. [14] used two different internet relay chat (IRC) datasets containing homogeneous and heterogeneous topics separately. Traditional chi-squared distance and Kullback-Leibler 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. [15] used IRC records of 50 users, each of whom entered 50 chat messages. The re-centered 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, an accuracy of up to 55% was achieved.

Lexical Features	**Character Features**
-total # of words	-total # of characters
-total # of unique words	-ratio of alphabetic chars.
-ratio of short words	-ratio of upper case letters
-mean word length	-ratio of digit characters
-mean sentence length	-ratio of white space chars.
-mean paragraph length	-ratio of punctuation chars.
-ratio of distinct words	-ratio of distinct chars.
-# of hapax legomena	-ratio of emoticons
-# of hapax dislegomena	-ratio of char. repetition
-word n-grams	-character n-grams
-skip-grams	-vowel combination
-word frequencies	-vowel permutation
-# of words of each length	-compression methods
Syntactic Features	**Semantic Features**
-freq. of function words	
-freq. of punctuation marks	
-part of speech (POS) tags	-synonyms of words
-total # of lines	-hypernyms of words
-total # of sentences	-semantic dep. graphs
-total # of paragraphs	-latent semantic analysis
-# of sentences per paragraph	-systemic func. grammar
-# of words per paragraph	
-# of characters per paragraph	
-ratio of spelling errors	

Table 2. Commonly used features for authorship recognition.

If we consider simplicity and language independence as primary factors, character based features are expected to perform better. Especially, the character n-gram representation has been used as one of the most effective measures of authorship attribution [16, 17]. On the other hand, lexical, syntactic and semantic features have some advantages over each other. For example, superiority of syntactic and semantic features depends on the idea that authors tend to unconsciously use similar patterns, and some language-specific NLP tools (such as a POS tagger, stemmer, spell checker) may be required for extracting syntactic and semantic patterns.

Work	Approach	Data
Zhao et al.'06 [7]	POS Tags, Kullback-Leibler Divergence,	Associated Press
Frantzeskou et al.'07 [8]	Byte Level N-gram, Source Code Author Profile	Author Data Set
Kešelj et al.'08 [9]	Character Level N-gram, Similarity Measure	Author Data Set
Koppel et al.'11 [10]	Character N-gram, Naive Similarity Method	Blog Data Set
Layton et al.'12 [11]	Character N-gram, Recentered Local Profile	AAAC
Savoy'12 [12]	Word Tokens, Z-Score	GH Corpus/ *La Stampa* Corpus
Seidman'13 [13]	Character N-gram, General Impostors	PAN at CLEF'13
Inches et al.'13 [14]	Mutual Word Influence, KLD, Chi-Square	IRC Logs
Ali et al.'14 [3]	Byte Level N-gram, TF-ITF, KNN	Chat Bot Corpus
Šarkute and Utka'15 [5]	Character N-gram, Bag of Words, Naive Bayes, SVM	Formal Speech Corpus

Table 1. Approaches for text authorship recognition.

2.2. Analysis of agglutinative languages

The bulk of authorship analysis approaches in the literature focus on English language, which is weakly inflected. Despite the fact that research on highly inflected languages like Greek and Sanskrit or fusional languages like German may be useful in order to understand language dependent approaches on chat biometrics to some extent [5, 9, 18], agglutinative languages differ from them by a complex word structure, which is formed by stringing together morphemes without changing them in spelling or phonetics. An example of such a language is Turkish, where for example, the word *evlerinizden*, or "from your houses", consists of the morphemes, *ev-ler-iniz-den* with the translation of *house-plural-your-from*.

There are some prior studies on author attribution in Turkish. Tufan et al. used style markers as features, on a gallery of 20 authors [19]. Amasyali et al. used n-gram model in text categorization for author, genre and gender classification [20]. Both studies used corpora collected from newspaper articles, which are written in formal Turkish.In another study, a chat mining framework was tested on a Turkish dataset containing peer-to-peer text messages [21]. 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 in style. However, term-based features achieved better results compared to style-based features on a 100-author problem.

3. METHODOLOGY

3.1. COPA Multiparty Chat Database

In this study, we have used the proprietary CCSoft Okey Player Abuse (COPA) Database, consisting of demographics, statistics, game records, interactions and complaints of thousands of players [2]. The database is acquired from a commercial 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. In the mentioned period, a total of 800,000 Okey games were recorded along with the player interactions in the chat area and the dataset contains chat inputs from more than 30,000 user accounts.

The database is particular in that messages are always written in a multiparticipant fashion (there are always four players in a game); they are unedited (except for a blacklist 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 modeling took place. We sorted chat participants according to the number of unique words used by each, and eliminated participants who had vocabulary sizes less then 100 unique words. This is a very coarse pre-processing, but people with very limited vocabulary might be easier to identify, and might positively bias the results. The remaining users are sorted in decreasing order according to number of active chat sessions, and the most active users are selected for building a chat biometrics benchmark database. With 978 users, this database is one order of magnitude bigger than the most relevant work from the literature. Table 3 describes the properties of the final database. Since we planned to use 5-fold cross validation, as well as to assess the effect of the number of chat entries per user, we required that at least 5 chat sessions per user should exist for each of five folds.

3.2. The proposed approach

We adopt the re-centered local profile (RLP) approach, proposed by Layton et al. [11], which uses a language profile in the calculation of a distance between an author and a document:

$$d(f_1, f_2) = \sum_{n \in profile} [f_1(n) - P(n)].[f_2(n) - P(n)]$$

where f_1 and f_2 are author/document profiles to be compared and P is the language profile, which is extracted from the entire training set as an approximation to the absolute language

Corpus Characteristics	**Value**
# of users	978
# of chat sessions per user	261
# of chat returns per user	3,251
# of unique words per user	2,375
# of words per user	10,933
# of letters per user	39,494
# of capital letters per user	149
# of emoticons per user	288
# of digits per user	162
# of punctuations per user	679

Table 3. Statistics of the chat biometrics subset of the COPA database.

profile. If we normalize profiles by using absolute distance of variation between each profile, the following equation is obtained:

$$d(f_1, f_2) = \sum_{n \in profile} \frac{[f_1(n) - P(n)].[f_2(n) - P(n)]}{\| f_1(n) - P(n) \| . \| f_2(n) - P(n) \|}$$

Because of the flexibility inherent in natural languages, extracting the absolute profile of a language is impossible. For this reason, all the normalized author profiles in the training set are combined to extract a standardized language profile.

The common n-grams (CNG) method proposed by Kešelj et al. [9] uses the relative distance between two documents (or author profiles), and serves as a basis for RLP. However, the most noticeable difference is that RLP measures the profile similarity according to most distinctive features, rather than the most frequently used features, using the standardized language profile approach described above.

For each entry, we have replaced capital letters with small cases. Then, $N \times N$ sparse bi-gram matrices for each user are calculated. We tested N equal to 32 and 66 (See Table 4) based on most common characters used in the COPA database. In addition to the RLP approach, we tested Cosine Similarity (CS) based identification.

	Characters
2-gram of 32 char	abcçdefgğhıijklmnoöprsştuüqwxyz
2-gram of 66 char	abcçdefgğhıijklmnoöprsştuüqwxyz 1234567890 ”!’^+%&/()=?_*-<>—@:.,;‘

Table 4. Selected characters for 2-gram feature matrix

We did not check the data for proper names, and due to the anonymized nature of the data set, we do not have access to the names of the users. However, if a person consistently uses their proper name (highly unusual), this may simplify subsequent matching. It is not possible to test for this directly, as the n-gram features do not preserve words. While we believe that proper names do not have a significant effect, as a future work, we consider removal of proper names using an external name database.

3.3. Experimental Protocol

As stated in Section 3.1, 978 active users (from among more than 30,000 users) were selected in order to set up an experimental protocol. Accumulated chat records of a user in one session were used as an instance for that user in the corpus. 5-fold cross validation was used.

To assess the number of characters and the number of words needed to identify a person, we performed two different sets of experiments. The entire set of users were labeled as SET 1. A minimum number of 140 characters per session was required in SET 2, which subsequently has only 617 users.

During testing, we have systematically increased the number of instance per test user from 1 to 100 in order to understand how performance changes with respect to the number of samples per user.

We describe the experiments as abbreviations in the tables, formed of identification method (CS or RLP), extracted n-gram features (32 or 66), and whether a minimum number of characters was required per session or not (GT140 for required and GT0 for not). For instance, RLP/66/GT140 means that re-centered local profile is used as identification method, features are extracted as 2-gram character matrix of size 66 × 66, and chat sessions in SET 2 were used, with at least 140 characters per session.

3.4. Turkish Normalization

A small part of the raw data on COPA was normalized by using the web API of the Turkish NLP tool [22], whereby intentional or accidental misspellings were replaced with correct forms. Since Turkish has flexible sentence structure, as well as agglutinative word forms, the normalization affects the identification performance significantly. For instance the raw sentence “*bütttttttüüüüünnnn insnlar e$it dogaaarr*” (“all people are born equal”) is normalized as “*bütün insanlar eşit doğar*”. Normalization changes the distribution of the n-gram features. In the next section, we report results with both raw and normalized test data. For the latter, the normalized version of training data was used.

4. RESULTS

In the first experiment, RLP and CS measures are used with the 2-gram character matrix size of 66 × 66 for each user. Given sufficient data for processing each user (results on SET 2), RLP and CS do not give significantly different results, as shown on Table 5. We tested for significance of differences with paired t-tests between the identification results. Conversely, if all chat sessions are taken into consideration, a significant difference was observed between CS and RLP, with RLP being the more accurate approach. Increasing the gallery size has a more detrimental effect on CS compared to RLP, but we should remember that the added users (i.e. those that are present in SET 1 but not in SET 2) are the ones with shorter utterances.

A second experiment was conducted to inspect the effect of the number of chat returns (entries) of a user per test case. *SET 1* was used, and *RLP/66/GT0* is the protocol for this experiment. The number of entries per user was systematically varied between 1 and 100, and an accuracy of 77.5% was obtained with 100 entries. The curve stays flat after the 92nd entry, suggesting that adding much more test data may not result in obtaining better results. Figure 1 illustrates the relationship between the number of chat entries and the rank-1 identification rate.

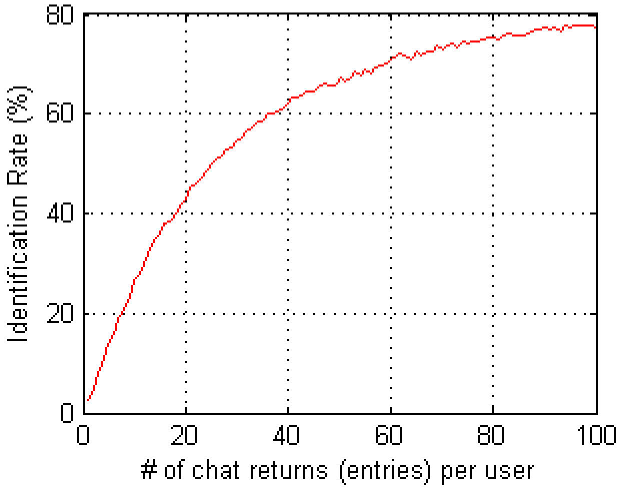

Fig. 1. Change of identification rate for test type *RLP/66/GT0* on *SET 1* while increasing the number of chat entries per user from 1 to 100 during testing.

One of the issues we investigate with this study is how text normalization impacts author identification from Turkish chat records. Our results show that raw chat data is more distinctive than normalized chat data, since intentional misspellings or unconscious typos are some of the most important features for identification. Normalization of text causes loss of these distinguishing features. The impact on the results is evident in Table 6, which reports the raw and normalized versions of each test setting. By performing a paired t-test, we also confirmed that the difference is statistically significant with $p < 0.0001$. We used a small set of users for the normalization experiments, and since the effect was very clear, we did not perform normalization on the entire set of users.

5. CONCLUSIONS

We have contrasted several approaches proposed for authorship identification on the problem of chat biometrics, and performed tests with a large database of multiparty chat records in Turkish which is available upon request for academic purposes. While normalization is a standard step in text processing for agglutinative languages, we illustrated that it results in accuracy loss, much like over-aggressive registration for face recognition. Our results show that it is possible to obtain around 75% rank-1 accuracy for a gallery size of 978.

6. REFERENCES

[1] David C Uthus and David W Aha, “Multiparticipant chat analysis: A survey,” *Artificial Intelligence*, vol. 199, pp. 106–121, 2013.

[2] Albert Ali Salah Koray Balci, “Automatic analysis and identification of verbal aggression and abusive behaviors for online social games,” *Computers in Human Behavior*, vol. 53, pp. 517 – 526, 2015.

[3] Nawaf Ali, Matthew Price, and Roman Yampolskiy, “BLN-Gram-TF-ITF as a new feature for authorship identification,” in *Academy of Science and Engineering (ASE) BIGDATA/SOCIALCOM/CYBERSECURITY Conference*, 2014.

[4] Efstathios Stamatatos, “A survey of modern authorship attribution methods,” *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[5] Ligita Šarkute and Andrius Utka, “The effect of author set size in authorship attribution for Lithuanian,” in *Nordic Conference of Computational Linguistics NODALIDA 2015*, 2015, p. 87.

[6] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang, “A framework for authorship identification of online messages: Writing-style features and classification techniques,” *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[7] Ying Zhao, Justin Zobel, and Phil Vines, “Using relative entropy for authorship attribution,” in *Information Retrieval Technology*, pp. 92–105. Springer, 2006.

[8] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, Carole E Chaski, and Blake Stephen Howald, “Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method,” *International Journal of Digital Evidence*, vol. 6, no. 1, pp. 1–18, 2007.

Database	**Test Type**	**Scores (%) with # of Chat Sessions**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
SET 1	RLP/66/GT0	17.73	29.41	41.84	50.92	57.46	62.84	66.93	71.29	73.07	75.66
	CS/66/GT0	16.20	27.42	39.35	48.90	54.97	59.75	64.72	67.93	71.31	74.40
SET 2	RLP/66/GT140	32.22	54.13	66.22	74.39	78.61					
	CS/66/GT140	31.53	54.32	66.00	74.72	79.06					

Table 5. Comparison of character threshold effect on models. The RLP method is significantly better than CS for the first set (t-test, $p < 0.0001$), however, the second set does not show a significant difference. Short utterances contain relevant information, which gets lost in the second set.

Test Type	**Test Data**	**Scores (%) with # of Chat Sessions**									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
CS/66/GT0	Raw	33.25	54.22	65.78	72.29	80.00	82.40	85.06	92.05	91.57	93.97
	Normalized	24.10	37.83	50.56	58.80	63.37	69.40	72.53	80.48	78.55	80.72
RLP/66/GT0	Raw	33.01	55.18	66.75	71.81	82.89	82.65	87.23	92.53	93.01	93.98
	Normalized	23.61	38.07	53.25	59.28	62.41	67.47	71.33	78.80	81.45	80.48

Table 6. Performance comparison of raw and normalized data of 83 chat users. The raw data produces significantly better results (t-test, $p < 0.0001$) compared to normalized data. This shows the textual errors are revealing in determining identity.

[9] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the conference pacific association for computational linguistics, PACLING*, 2003, vol. 3, pp. 255–264.

[10] Moshe Koppel, Jonathan Schler, and Shlomo Argamon, "Authorship attribution in the wild," *Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.

[11] Robert Layton, Paul Watters, and Richard Dazeley, "Recentred local profiles for authorship attribution," *Natural Language Engineering*, vol. 18, no. 03, pp. 293–312, 2012.

[12] Jacques Savoy, "Authorship attribution based on specific vocabulary," *ACM Transactions on Information Systems (TOIS)*, vol. 30, no. 2, pp. 12, 2012.

[13] Shachar Seidman, "Authorship verification using the impostors method," in *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*, 2013.

[14] Giacomo Inches, Matthew Harvey, and Fabio Crestani, "Finding participants in a chat: Authorship attribution for conversational documents," in *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 2013, pp. 272–279.

[15] Richard Layton, Stephen McCombie, and Paul Watters, "Authorship attribution of IRC messages using inverse author frequency," in *Cybercrime and Trustworthy Computing Workshop (CTC), 2012 Third*. IEEE, 2012, pp. 7–13.

[16] Patrick Juola, "Authorship attribution for electronic documents," in *Advances in digital forensics II*, pp. 119–130. Springer, 2006.

[17] Conrad Sanderson and Simon Guenter, "On authorship attribution via Markov chains and sequence kernels," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, vol. 3, pp. 437–440.

[18] George K Mikros and Kostas Perifanos, "Authorship attribution in Greek tweets using author's multilevel n-gram profiles.," in *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[19] Taş Tufan and Abdul Kadir Görür, "Author identification for Turkish texts," *Cankaya University Journal of Arts and Sciences*, vol. 1, no. 7, 2007.

[20] M Fatih Amasyalı and Banu Diri, "Automatic Turkish text categorization in terms of author, genre and gender," in *Natural Language Processing and Information Systems*, pp. 221–226. Springer, 2006.

[21] Tayfun Kucukyilmaz, B Barla Cambazoglu, Cevdet Aykanat, and Fazli Can, "Chat mining: Predicting user and message attributes in computer-mediated communication," *Information Processing & Management*, vol. 44, no. 4, pp. 1448–1466, 2008.

[22] Gülsen Eryigit, "ITU Turkish NLP web service," *EACL 2014*, p. 1, 2014.