Attribution of authorship in instant messaging software applications, based on the similarity measure of the stylometric features' vector

M. MAZUREK, M. ROMANIUK

marcin.mazurek@wat.edu.pl, mateusz.romaniuk@student.wat.edu.pl

Military University of Technology, Faculty of Cybernetics
Kaliskiego 2, 00-908 Warsaw, Poland

This paper describes the issue of authorship attribution based on the content of conversations originating from instant messaging software applications. The results presented in the paper refer to the corpus of conversations conducted in Polish. On the basis of a standardised model of the corpus of conversations, stylometric features were extracted, which were divided into four groups: word and message length distributions, character frequencies, tf-idf matrix and features extracted on the basis of turns (conversational features). The vectors of users' stylometric features were compared in pairs by using Euclidean, cosine and Manhattan metrics. CMC curves were used to analyse the significance of the feature groups and the effectiveness of the metrics for identifying similar speech styles. The best results were obtained by the group of features being the tf-idf matrix compared with the use of cosine distance and the group of features extracted on the basis of turns compared with the use of the Manhattan metric.

Keywords: authorship attribution, stylometry, CMC, turn, instant messaging software application, conversation

DOI: 10.5604/01.3001.0015.2735

1. Introduction

Attribution of authorship involves selecting, from a group of people, one person, who is the most likely author of a text, given a collection of texts with known authorship. One of its main uses was to identify the author of literary texts and documents in cases where there was doubt about the person who wrote them or the person was unknown.

The identification of the author of a statement plays a special role when it aims to discover the identity of a person who wishes to remain anonymous. This type of problem occurs in communications conducted by using instant messaging software applications. Instant messaging software applications have many features that favour its use by criminals. Fraudsters, criminal groups or terrorist organizations can use these tools to communicate. Instant messaging software applications allow the creation of virtual identities that are not verified, making it possible to create multiple fake accounts or impersonate other people. The information you provide when creating an account is not verified and may be false, so do not rely on it when trying to identify the author of a message. For this reason, it is necessary to use other information to identify the author of a message.

The paper [7] identifies author profiling techniques as useful in countering these threats and in ongoing investigations.

Instant messaging software applications have completely changed the way we have conversations. Sent messages are getting shorter, contain less and less text, and at the same time they become more diverse due to the use of abbreviations or emoticons by the interlocutors. The rules of formulating statements are constantly being abandoned in favour of the quick transmission of information. The exchange of messages between the interlocutors is very dynamic, which makes them resemble more conversations (oral-CSMM_11_12_2020 utterances) than other forms of written communication (e-mails, traditional letters). Therefore, they should be analysed in a different way and a set of features allowing to capture the characteristic features of such conversations should be found. As presented in [3], [4] and [6], the analysis of texts of this type is still at an early stage of development, in addition, no publications showing the effectiveness of this type of methods for the Polish language have been found.

2. Attribution of authorship in instant messaging software applications

The use of stylometric analysis for authorship attribution in instant messaging involves the construction of a feature vector describing the style of utterance. As the style of utterance is unique and difficult to imitate without arousing suspicions[1], the similarity of vectors should indicate the possibility of the same person hiding behind the same pseudonym.

The sets of characteristics and patterns extracted from the messages sent by the user, and therefore the style of utterance of the user can be compared with the style of utterance of other users, which could be applied to the following issues: identification of persons hiding under different virtual identities – identification of accounts belonging to one person, identification of accounts of one person in various social networks. In addition, comparing the utterance style of a single user at intervals would allow detection of cases where more than one person is using the same account. Introducing automatic verification of utterance style could be another way to confirm the authenticity of the interviewee.

In [1], the issue of authorship attribution for messages coming from online forums is presented. The problem was considered for English and Arabic. The features describing the user's writing style were divided into four categories: lexical, syntactic, structural and content-specific, where two lexical subgroups were distinguished from the first group: word-based subgroup and character-based subgroup.

The paper [5] addresses the problem of attribution of authorship for messages in Italian from Skype[2]. The paper emphasizes that IM conversations have a lot in common with oral conversations, but the current approaches to the attribution of authorship problem ignore this fact, while it is crucial and distinguishes conversations from instant messaging software applications from other forms of communication, such as e-mail, mail or forum entries. An approach has been proposed to address the above problem. The concept of the user *turn* in instant messaging conversations is distinguished. It has been determined that a conversation consists of a sequence of consecutive users' turns, where a turn is a sequence of words and characters sent by one conversation participant, uninterrupted by another conversation participant. A turn may contain new row characters, and so may consist of more than one message. A new group of features, intended to reflect the nature of conversation, was then proposed and named the *conversational features* category. This group was distinguished by the following features describing the user's participation in the conversation: turn duration, writing speed, number of messages and mimicry degree.

The paper [2] considered the issue of authorship attribution for English language posts from Twitter – the Internet social network. A profile-based approach was used for authorship attribution – after pre-cleaning the text, all user posts were combined into one long text string, from which features were then extracted. N-grams and a set of stylometric features were used to solve the authorship identification problem. In the case of n-grams, extraction was performed at both word level and character level. The length of n-grams between 2 and 4 was tested, citing sources [8] and [9], which found that n-grams with a length greater than 5 did not positively affect the recognition of the user's utterance style. Moreover, they can be used successfully even for short texts. The stylometric features use a group of lexical features – 8 features, structural features – 8 features, and user-specific features i.e. idiosyncratic features. The last group contained two features altogether – misspelled words and abbreviations and colloquial words (abbreviations / slang).

[1] We are talking about instances where a user attempts to change his/her utterance style during a conversation when he/she is actively participating in the conversation, without using algorithms and models constructed for this purpose.

[2] https://www.skype.com/pl/

3. Conversation corpus model

The results presented below were obtained from the analysis of a corpus consisting of 6348 conversations held by 391 users, consisting of a total of 268 thousand messages.

It should be noted that the users' contribution to the content of the corpus was very uneven. The content generated by the 6 most active users accounts for almost 50% of the corpus content, and the 25 most active users generated over 75% of the content. Figure 1 shows the amount of content produced for the 100 most active users and the cumulative amount of content produced. The first user in the ranking sent 23.4% of all characters, and the first two users – 30.1%.

All stylometric features that will be extracted were divided into 4 groups:

- frequency of occurrence of characters
- message length and word length
- conversational features
- frequency of occurrence of words in user messages against the entire corpus of messages (tfidf).

The first group includes features related to the frequency of occurrence of characters – a total of 77 features. The frequency was determined for the following groups of characters:

- the characters of Polish alphabet
- digits / numerals
- special characters

Figure 2 shows frequency of occurrence of characters in the entire corpus.

The second group includes the features describing the message length distribution (number of words in the message) – 38 features and the word length distribution (number of characters in a word) – 14 features.

Tab. 1. Features extracted from the message

Characteristics	How the feature is obtained [number of ranges]
number of words in a message	distribution [38]
the number of characters in a word	distribution [14]

Figure 3 shows the word length distribution in the corpus. The distribution is right-skewed, dominated by short words. The average value is 4.24, while for 73% of the turns the average word length was between 3 and 5. For reference, the average word length for the Polish language measured by the number of characters in a word is 6 [11].

The third group includes conversational features, extracted on the basis of the users' turns, which was described in [5]. This group includes a total of 116 features (Table 2). In the case of features that are histograms, exponential histograms have been used (the width of each successive histogram interval increases exponentially) because, as shown in [5], for the task under consideration, histograms constructed in this way yield better results.

Tab. 2. Features extracted on the basis of the turns

Characteristics	The method of obtaining the feature [number of ranges]
number of words	distribution [20]
number of emoticons	distribution [4]
number of exclamation marks	distribution [3]
number of question characters	distribution [5]
number of characters	distribution [20]
number of ellipses	distribution [4]
number of uppercase letters	distribution [5]
duration [s]	distribution [19]
number of messages	distribution [20]
number of emoticons per one word	median
number of emoticons per one character	median
average word length	distribution [20]
number of capital letters per word	median
number of characters per second	median
number of words per second	median
mimicry index	distribution [20]

Figure 4 shows the distribution of the duration of a turn in the entire corpus.

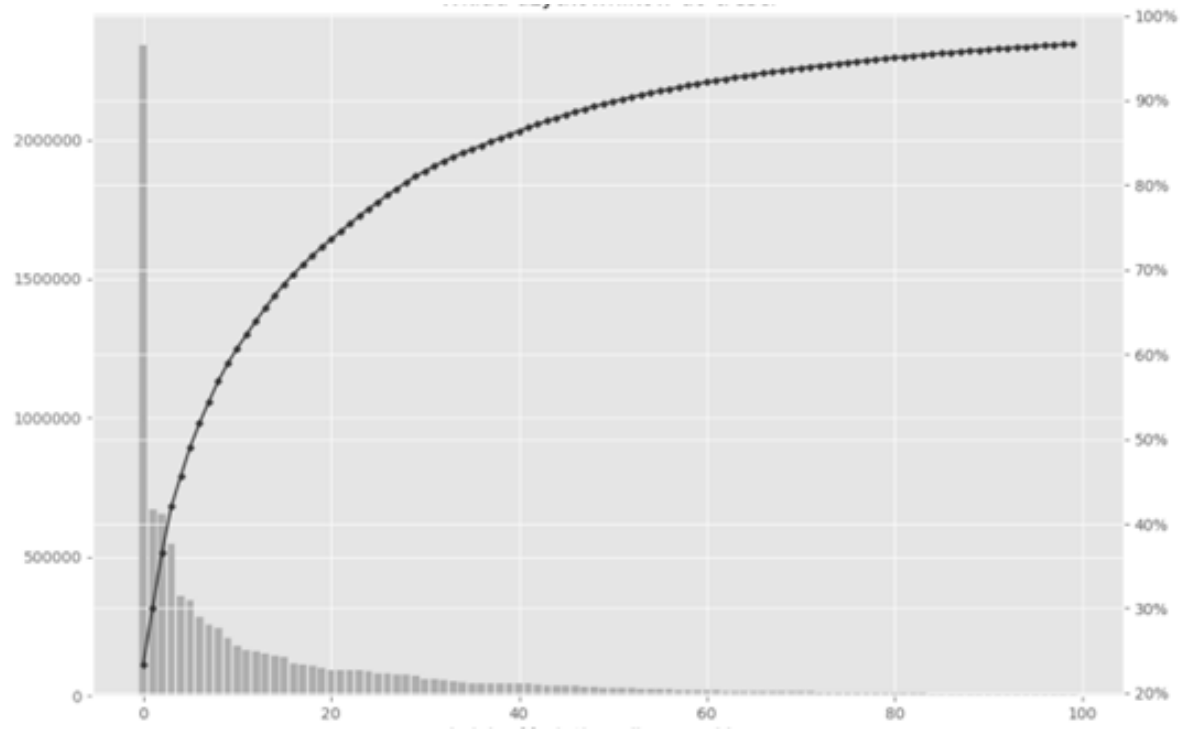

Fig. 1. Distribution of users to the corpus content

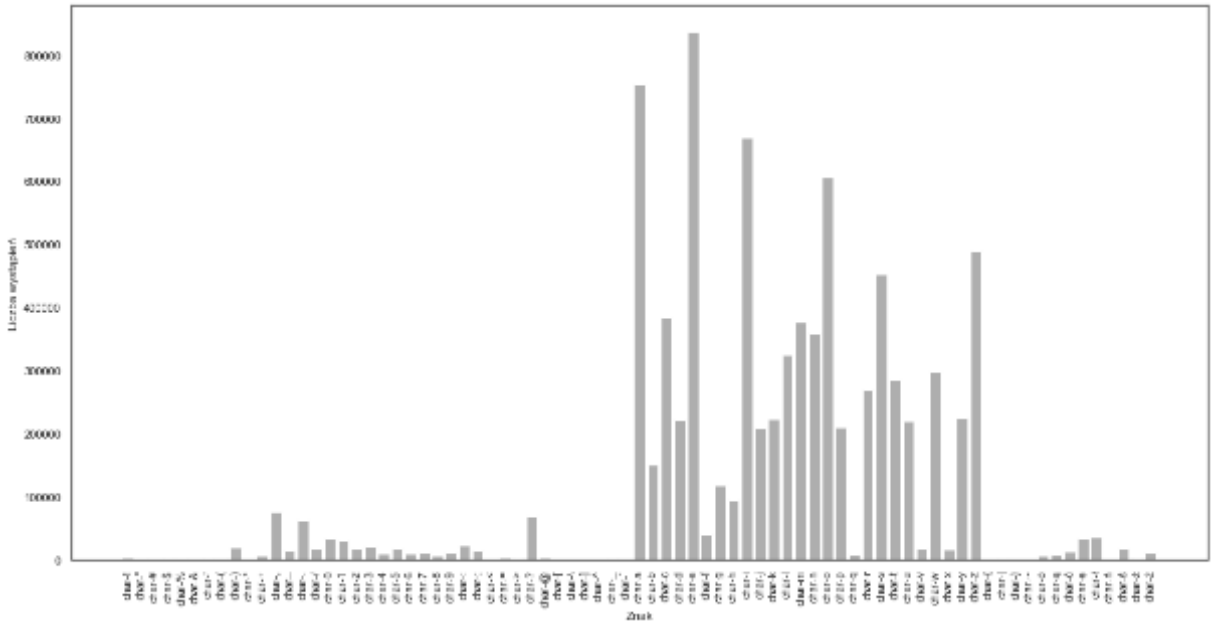

Fig. 2. Distribution of the frequency of occurrence of alphabet characters

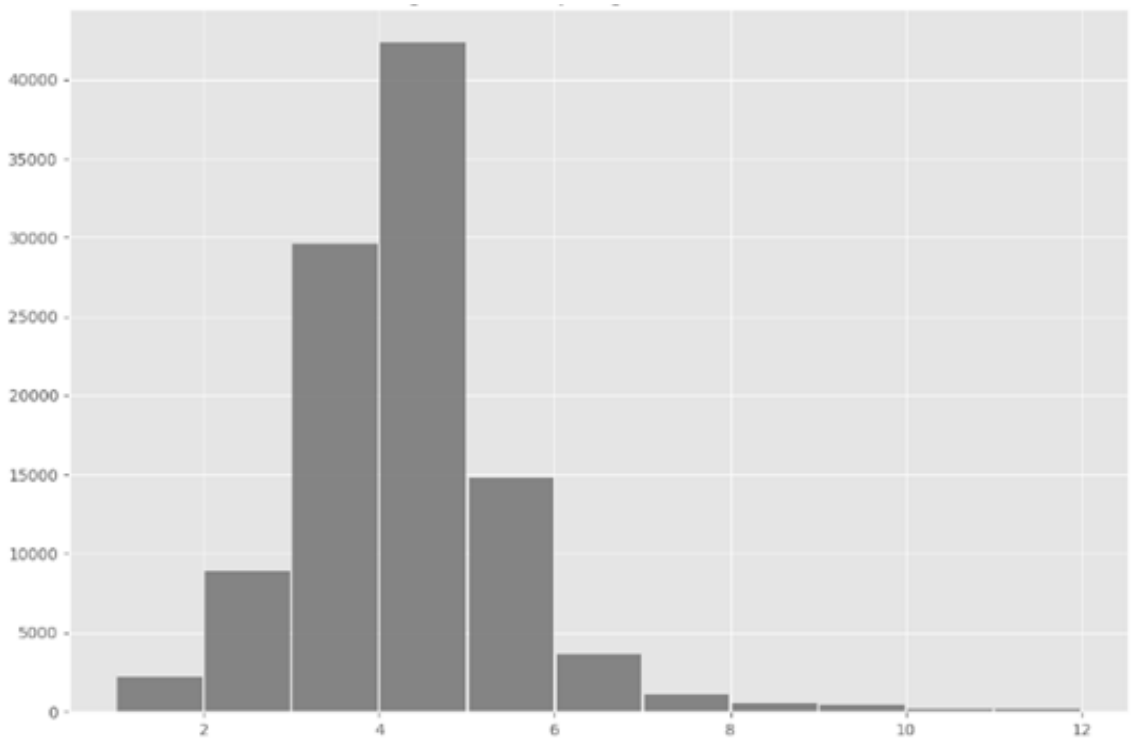

Fig. 3. Distribution of the average word length in a turn

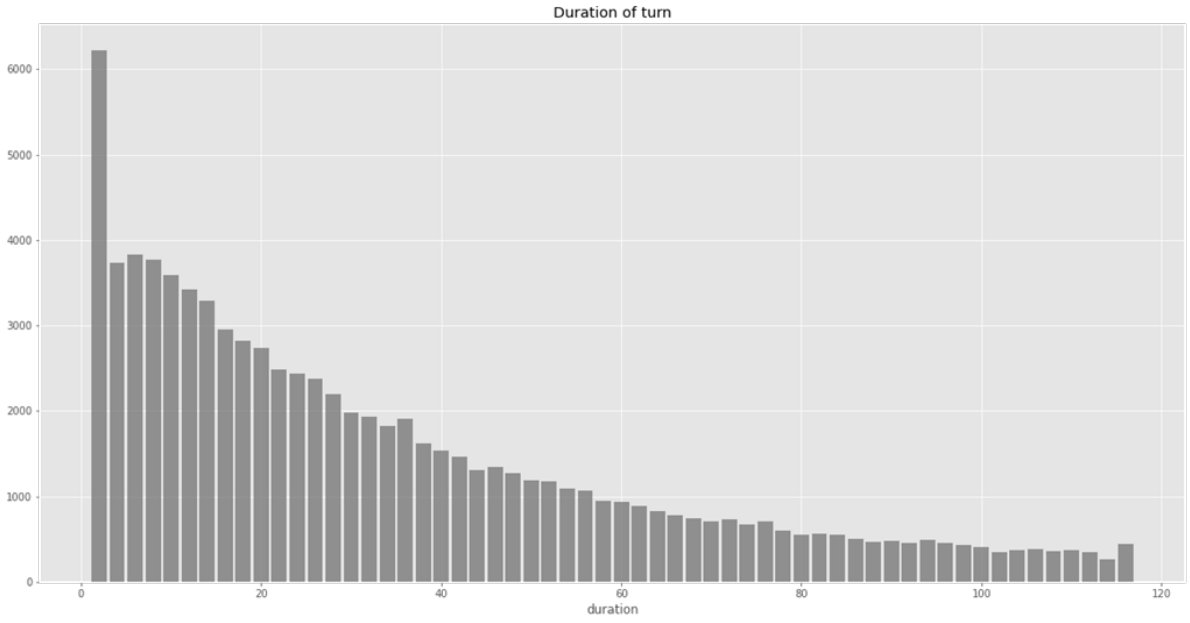

Fig. 4. Distribution of the length of the turn duration

The fourth and last group of features is the frequency of words in the user's messages against the entire conversation corpus, represented by the tf-idf (term frequency – inverse document frequency) matrix. A detailed description of the tf-idf algorithm with modifications is presented, among others, in [10]. It enables you to find keywords that make you stand out from the rest of your interlocutors. It should be mentioned that the last group of features should be used with limited confidence when they have been extracted from a heterogeneous conversation corpus containing conversations on multiple topics. It may be that a model learned from these features solves the text classification task based on its topics rather than the authorship attribution task.

Features can be extracted at different levels of granularity. Such a division will enable the analysis of special cases in the authorship attribution task. 3 levels are distinguished:

- user
- user in conversation with another user
- user in one conversation.

In the first case, stylometric features are extracted from all utterances of a user. In the set of stylometric features, each user is one row. This approach can be compared to the author profile-based approach presented in [2]. In the second case, one row in the set of stylometric features is constructed from all user's messages sent to another specific user. The number of rows in the set of stylometric features relating to a single user is therefore equal to the number of users with whom the user has had conversations. This approach will identify cases where the user speaks differently depending on the co--interlocutor. In the third approach, each row in the set of stylometric features pertaining to a user, describes one of his/her conversations. In this approach, the number of rows in the set of stylometric features pertaining to one user, is equal to the number of all his/her conversations. The extraction of features at this level, in addition to its application as in the second case, will also enable the detection of deviations from the pattern in the style of utterance of the user, which may indicate that one account is used by more than one person. This can be caused either by you sharing your account with another person or by an unauthorized third party taking over your account. Another application could be to analyse how a user's utterance style changes over time.

4. Results

The Cumulative Match Characteristic (CMC) curve was used to compare methods for determining the similarity of the users' stylometric feature vector. For its determination, the turns of each user were divided into two disjoint sets. For each user, a vector representation of the style was determined in accordance with the model described in the previous chapter, and then compared with the vector representations of the styles obtained for the second subset of the turns. In addition, two auxiliary measures were defined to compare the results. The first measure represents the normalized number of cases (probability) for which the corresponding group (match) was indicated in the first position. This will be referred to as *p1*. It should be used to find the method that indicates the match in the first position in as many cases as possible.

Tab. 3. Probability of indicating an equivalent in the first position (p1)

Group of features	Metrics		
	cosinus	Euclid	Manh
The tf-idf matrix	0.8642	0.8642	0.1296
Frequency of occurrence of characters	0.6914	0.6543	0.6728
Message and word length distribution	0.5185	0.5247	0.5741
Features extracted on the basis of the turns	0.6296	0.4568	0.7037

The second measure (worst-case coefficient) tells you what the least populous part of the set should be taken so that in each case the match is in this set. It thus corresponds to the position on the CMC graph for which the probability value reaches 1.

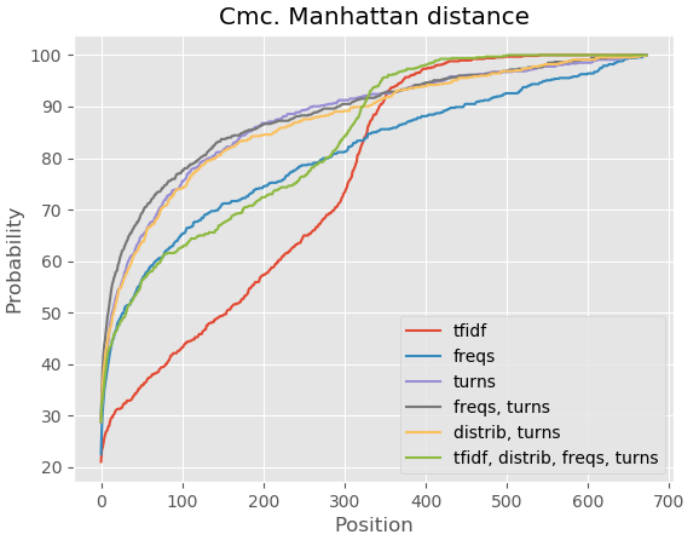

Fig. 5. CMC for Manhattan metric for different components of the feature vector

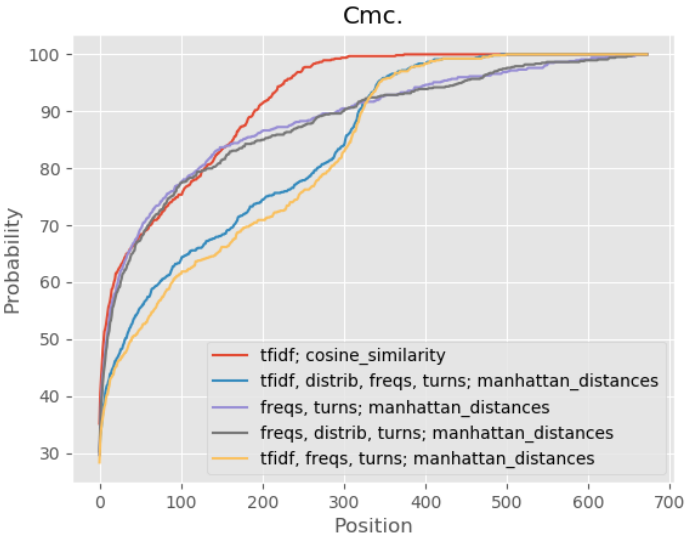

Fig. 6. CMC for selected metrics and feature vector components

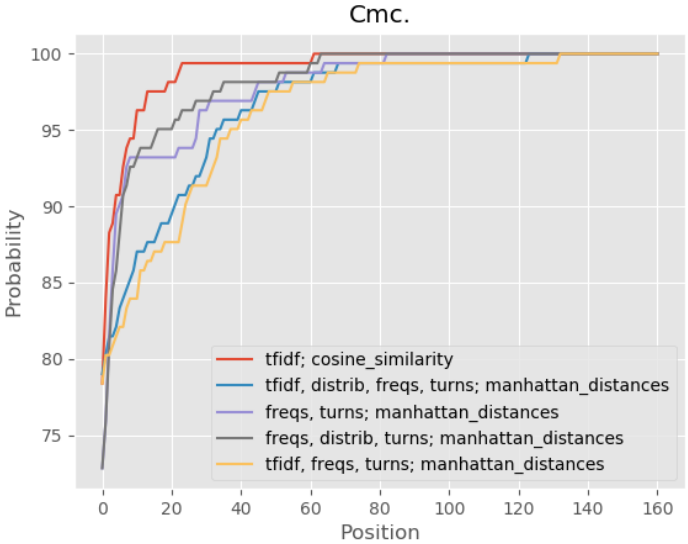

Fig. 7. CMC for selected metrics and feature vector components when restricted to users with large contributions to the corpus

This measure, like the previous one, is standardised. A value of 0 (the best possible case) means that for each user the match was found in the first place (the vector corresponding to the stylometric features of the user's turn group is located closest to the vector corresponding to the stylometric features of the other turn group of the same user).

Tab. 4. Worst-case coefficient

Group of features	Metrics		
	cosinus	Euclid	Manh
The tf-idf matrix	0.9563	0.9563	0.9063
Frequency of occurrence of characters	0.9625	0.9625	0.9625
Message and word length distribution	0.6000	0.6188	0.7688
Features extracted on the basis of the turns	0.8563	0.8125	0.5813

The adequacy of the proposed vector representation of style is strongly dependent on the text sample we have. The reliability of the proposed measures for assessing style similarity is higher the more conversations the user has had. The above statement was confirmed by conducting an analysis on a set of conversations carried out by 80 most active users (each of them sent at least 100 turns). The resulting CMC curve is shown in Figure 7. The curve shows that by indicating 10 best-matched vector representations for each user, in over 85% of cases we will find (in them) a representation of the turns of the same user's (but built on a different set of utterances).

5. Conclusion

This paper presents an unsupervised authorship attribution method involving the comparison of vectors of stylometric features extracted from conversations. The distance metrics were examined: Euclidean, Cosine and Manhattan. The best results were obtained by the group of features being the tf-idf matrix compared with the use of cosine distance and the group of features extracted on the basis of turns compared with the use of the Manhattan metric.

Data exploratory analysis showed that the share of individual users in the production of conversation content was very diversified – the six most active users generated almost 50% of the content of the entire conversation corpus. The analysis also showed that the conversations conducted with the use of instant messaging are dynamic. The turns of users usually consisted of one or two messages, 80% of turns contained no more than 22 words, and in every 10th turn there was only one word. Users deviated from the rules of formulating utterances in favour of the speed of information delivery. The average word length in the turn was lower than the average word length for the Polish language. Many users neglected to use diacritical characters.

The proposed feature vectors can be used to identify similarities in utterance style and thus treated as a user-unique "write-print" for pairing conversational identities. An important limitation of the application of the presented methods may be the insufficient length of the text coming from the user, although already with 100 turns the results obtained can be considered satisfactory.

6. References

[1] Abbasi A., Hsinchun Ch., "Applying authorship analysis to extremist-group web forum messages", *IEEE Intelligent Systems*, No. 5, 67–75 (2005).

[2] Belvisi N.M.S., Muhammad N., Alonso-Fernandez F., "Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features", *Proc. 8th International Workshop on Biometrics and Forensics,* IWBF, Porto, Portugal, April 29–30, 2020, arXiv:2003.11545.

[3] Boenninghoff B., Hessler S., Kolossa D., Nickel R.M., "Explainable Authorship Verification in Social Media via Attention-based Similarity Learning", IEEE Big Data 2019, arXiv:1910.08144.

[4] Brocardo M. L., Traore I., Saad S., Woungang I., "Authorship Verification for Short Messages Using Stylometry", *Proc. of the IEEE Intl. Conference on Computer, Information and Telecommunication Systems* (*CITS 2013*), Piraeus-Athens, Greece, May 7–8, 2013.

[5] Cristani M., Roffo G., Segalin C., Bazzani L., Vinciarelli A., Murino V., "Conversationally-inspired stylometric

features for authorship attribution in instant messaging", *Proceedings of the 20th ACM International Conference on Multimedia*, Nara, Japan, 2012.
[6] Hai-Jew S., "A Light Stroll through Computational Stylometry and its Early Potential", https://scalar.usc.edu/works/c2c-digital-magazine-fall-winter-2016/a-light-stroll-through-computational-stylometry-and-its-early-potential [access: 11.02.2021].
[7] Orebaugh A., Kinser J., Allnutt J., "Visualizing Instant Messaging Author Writeprints for Forensic Analysis", *Annual ADFSL Conference on Digital Forensics, Security and Law*, 8, 2014, https://commons.erau.edu/adfsl/2014/thursday/8
[8] Houvardas J., Stamatatos E., "N-gram feature selection for authorship identification", in: *Artificial Intelligence: Methodology, Systems, and Applications*, J. Euzenat and J. Domingue (Eds.), 77–86, Springer Berlin Heidelberg, 2006.
[9] Wright D., "Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem", *International Journal of Corpus Linguistics*, Vol. 22, No. 2, 212–241 (2017).
[10] Santhanakumar M., Columbus C.C., "Various Improved TFIDF Schemes for Term Weighing in text Categorization: A Survey", *International Journal of Applied Engineering Research*, Vol. 10, No. 14, 11905–11910 (2015).
[11] Moździerz T., "Długość przeciętnego polskiego wyrazu w tekstach pisanych w świetle analizy korpusowej", *Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców*, Nr 27, 177–192 (2020), https://doi.org/10.18778/0860-6587.27.09.

Atrybucja autorstwa w komunikatorach internetowych na podstawie miary podobieństwa wektora cech stylometrycznych

M. MAZUREK, M. ROMANIUK

W artykule opisano zagadnienie atrybucji autorstwa na podstawie treści konwersacji pochodzących z komunikatorów internetowych. Zamieszczone w artykule wyniki odnoszą się do korpusu konwersacji prowadzonych w języku polskim. Na podstawie ustandaryzowanego modelu korpusu konwersacji wyodrębnione zostały cechy stylometryczne, które podzielono na cztery grupy tj.: rozkłady długości słowa i wiadomości, częstotliwości występowania znaków, macierz tf-idf oraz cechy wyodrębnione na podstawie tur (konwersacyjne). Wektory cech stylometrycznych użytkowników porównane zostały parami z wykorzystaniem metryk: euklidesowej, kosinusowej oraz Manhattan. Przy pomocy krzywych CMC przeanalizowano istotność grup cech oraz skuteczność metryk dla identyfikacji podobnych stylów wypowiedzi. Najlepsze rezultaty miała grupa cech będąca macierzą tf-idf porównywana z wykorzystaniem odległości kosinusowej oraz grupa cech wyodrębnionych na podstawie tur porównywana z wykorzystaniem metryki Manhattan.

Słowa kluczowe: atrybucja autorstwa, stylometria, CMC, tura, komunikator internetowy, konwersacja