Proceedings of the Federated Conference on Computer Science and Information Systems pp. 373–382
DOI: 10.15439/2017F461
ISSN 2300-5963 ACSIS, Vol. 11

Open Class Authorship Attribution of Lithuanian Internet Comments using One-Class Classifier

Algimantas Venčkauskas, Arnas Karpavičius
Department of Computer Science
Kaunas University of Technology
Studentu 50, LT-51368, Kaunas, Lithuania

E-mail: algimantas.venckauskas@ktu.lt

Robertas Damaševičius, Romas Marcinkevičius
Department of Software Engineering
Kaunas University of Technology
Studentu 50, LT-51368, Kaunas, Lithuania

E-mail: robertas.damasevicius@ktu.lt

Jurgita Kapočiūtė-Dzikienė
Faculty of Computer Science
Vytautas Magnus University
Vileikos 8, LT-44404, Kaunas, Lithuania

Christian Napoli
Department of Mathematics and Informatics
University of Catania
Viale A. Doria 6, 95125 Catania, Italy

***Abstract*—Internet can be misused by cyber criminals as a platform to conduct illegitimate activities (such as harassment, cyber bullying, and incitement of hate or violence) anonymously. As a result, authorship analysis of anonymous texts in Internet (such as emails, forum comments) has attracted significant attention in the digital forensic and text mining communities. The main problem is a large number of possible of authors, which hinders the effective identification of a true author. We interpret open class author attribution as a process of expert recommendation where the decision support system returns a list of suspected authors for further analysis by forensics experts rather than a single prediction result, thus reducing the scale of the problem. We describe the task formally and present algorithms for constructing the suspected author list. For evaluation we propose using a simple Winner-Takes-All (WTA) metric as well as a set of gain-discount model based metrics from the information retrieval domain (mean reciprocal rank, discounted cumulative gain and rank-biased precision). We also propose the List Precision (LP) metric as an extension of WTA for evaluating the usability of the suspected author list. For experiments, we use our own dataset of Internet comments in Lithuanian language and consider the use of language-specific (Lithuanian) lexical features together with general lexical features derived from English language. For classification we use one-class Support Vector Machine (SVM) classifier. The results of experiments show that the usability of open class author attribution can be improved considerably by using a set of language-specific lexical features together with general lexical features, while the proposed method can be used to reduce the number of suspected authors thus alleviating the work of forensic linguists.**

I. Introduction

THE Internet has become a critical enabling factor of economic and social transformations, affecting how governments, businesses and citizens interact and offering new, often unforeseen, ways of addressing challenges of sustainable development. Building trust in online services is essential to the continued growth and development of the Internet, while cybersecurity is vital for supporting sustainability and stability of the Internet. People need to have confidence that their data are secure, and networks and services they use are secure and reliable, while the societies and the state need to be sure that the tools of the Internet are not misused for criminal activities. The growth in extent and complexity of cybercrime combined with the lack of time and resources in addressing cybercrime, and the need to confront cybercrime in near real time raises a need to process available digital evidences on the Internet using computational intelligence techniques such as Natural Language Processing (NLP) [1].

Currently, web-based communication platforms and social sites (such as Facebook, Twitter, blogs, discussion forums, online knowledge portals, and chatting tools) allow users to publicly express their views and share information online. Information spread using such platforms and websites becomes readily available to a large number of people. A large audience of readers is a great medium for radical extremists to declare their views and try to influence public opinion, organize information attacks against individual groups of society or the whole countries. Public cyberspace and social networks can become channels of information to apply black public relations technologies for propaganda of violence and hate. The individuals who tend to spread ethnic, racial hatred, extremism and inciting war, or threatening public or national security often exploit the openness of social media by trying to hide behind nicknames or using other opportunities to stay anonymous. Establishing the authorship of an anonymous text based on the characteristics of the text only is a tedious and laborious task, which can be performed only by skilled forensics linguistics experts.

Manual work of experts who carry out monitoring is accurate, but ineffective: carried out in real time and round the clock, it would require having huge human resources in case of emergencies (information war, hybrid war, riots, armed conflict). Because of these limitations, currently forensic linguists are only asked to analyse texts of only a small number of authors (usually, a maximum of four or five). Any tools and methods that could help to reduce the amount of work and allow to expand the number of analyzed suspects in order to establish the true author (such as, e.g., by reducing the number

of suspected authors) are needed by the forensics community. The results of such automated (or software-generated) analysis can play a role in criminal investigations and trials [2].

The forensic analysis of electronic web-based texts for solving their authorship problem is called *authorship analysis* [3]. It involves analyzing the writing styles or stylometric features from the document content. As writing style varies from one author to another, the aim of *authorship attribution* is the correct classification of texts into classes based on the style of their authors. Besides *author identification* where the style of individual authors is examined, *author profiling* can also distinguish between categories of authors such as gender, age, or native language. *Authorship verification* checks whether a target document was written or not by a specific individual. In *authorship attribution*, the actual author is known to be included in the set of candidates (closed case) [4]. In *open class authorship attribution*, the analyzed text might not have been written by the candidate authors (open case). Given the examples of the writing of a single author the task is to determine if given texts were or were not written by this author [5], therefore, it provides a more realistic interpretation of the task since it approximates better what forensic linguistics experts do. Authorship attribution can be performed using stylometric techniques through the analysis of linguistic styles and writing characteristics of the authors [6]. Applications include email analysis [7] and spam filtering [8], computer forensics [9], plagiarism detection [10], cyberpredator identification in online chats [11], tagging of online texts [12] and news media analysis [13]. In all of these domains, the goal is to confirm or reject the authorship hypothesis for documents with respect to a set of candidate authors, given sample documents written by the considered authors. A close problem is plagiarism detection, where usually two texts are compared to find similarities between them [10]. The practice is also relevant for developing sustainable research and science. In many cases of plagiarism, the misconducting authors attempt to diffuse responsibility across many (perhaps innocent) co-authors [14]. So the question of establishing the true culprit is appropriate. As noted in [15], when dealing with more than twenty candidates, it is beneficial to identify a smaller subset of candidates using other scalable methods. The aim of this paper is 1) to propose a method of open class authorship attribution aimed at producing the list of suspected authors rather than a single prediction result, 2) to discuss the measures for evaluating the usability of the proposed method, and 3) to consider language-specific (we focus on the Lithuanian language) text features to improve the accuracy of authorship attribution.

The structure of the paper is as follows. We describe the problem formally in Section II. We discuss the language-specific text features in Section III. We describe the proposed method in Section IV. We discuss the evaluation metrics in Section V. Finally, the results are provided and discussed in Section VI, and conclusions are given in Section VII.

II. Problem of authorship attribution

The main element of authorship analysis process is text classification, i.e., the process of assigning predefined category labels to new documents. The formal description of our task is given below.

Let $t \in T$ be a text message, which belongs to a text space T. Let A be a finite set of authors: $A = \{a_1, a_2, ..., a_N\}$.

Let T^L be a training set and T^K be a testing set of text messages, containing instances I of text feature vectors $v \in V$ which belong to a feature space V (where each v corresponds to a text message t) with their appropriate class labels: $I^L = \langle v, a\rangle$, $I^K = \langle v, a\rangle$.

The text message t represented by the feature vector v is linked to exactly one author $a \in A$.

Let function ξ be a function that generates instances I from text messages t based on the feature space V : $\xi : T \times V \rightarrow I$.

Feature space V can be partitioned into a number of non-overlapping feature subspaces V_k as follows: $V = \bigcup_{1 \leq k \leq M} V_k$, $\bigcap_{1 \leq k \leq M} V_k = \emptyset$, here M is the number of features.

Let V_1 be a set of text features representing general text features used independently of text language, and $V_k, k \geq 2$ are sets of language-specific text features.

Let function γ be an authorship attribution function mapping a text message t to an ordered set of authors A', $\gamma : T \rightarrow A'$, where $A' = \langle A, r\rangle$ and r is a binary relation of authors a_i and a_j that is equal to 1, if a_i is more likely to be the author of the message than a_j, and 0, if otherwise. A is a sorted list with the most likely authors on top.

Authorship attribution of text consists of associating a real value p, $0 \leq p \leq 1$ to each pair $(t_j, a_i) \in T \times A$, where T is the set of text messages, A is the set of authors, and p reaches its maximum for a true author of the text.

Let Γ denote a supervised learning method, which given a set of instances I as the input, returns a learned mapping function γ as the output: $\Gamma : I \rightarrow \gamma$.

Let π be an evaluation metric (function) mapping from a testing set of texts T^K to a real value q, $0 \leq q \leq 1$ that evaluates the quality of author attribution: $\pi : T^K \rightarrow R$.

We attempt to find a best feature subspace V_k that given a text space T would maximize the authorship attribution function γ'. We define the best authorship attribution function with regard to V_k, $k \geq 2$ as the function $\gamma' = \max_{\gamma} \pi(\gamma(T^K))$, where $\gamma = \Gamma(\zeta(T^L, V_1 \bigcup V_k))$.

Further we employ the one-class learning approach for authorship attribution of language-specific texts, which has been introduced first by Koppel and Shler in [16]. One-class classifier defines a boundary around the target class that leaves out the outliers. The reference author is assigned to the target class and all other authors are attributed to the outlier class [17]. For each author, we have sample documents written by her/him. Sample documents for the author under consideration are considered as positive examples, whereas

sample documents for other authors are considered as negative examples. Features are extracted from documents and a classification model is built for the author. For a large number of features, Principal Component Analysis (PCA) can be used to derive a reduced number of 'effective' features, which retain most of information [18].

III. Language specific features

Establishing features that work as effective discriminators of texts is one of critical issues in research on authorship analysis. The problem so far is that most research in the area of authorship identification is focused on the English texts (with a few notable exceptions such as Greek [19], Portuguese [20], and Croatian [21], while applications for other languages usually focus on the application and adaptation of the text features and methods adopted from English (e.g., using n-grams [22]). In case of a language-specific discourse, two approaches are prevalent: one approach ignores the specifics of a national alphabet by transliterating language-specific alphabet letters to standard Latin or English alphabet letters. The other approach leaves language-specific letters in feature space for further analysis.

Typically, the same or a similar set of features are used and consequently the use of language-specific letters does not lead to significant improvement of author identification results. New language specific features (such as variable-length language-specific syllables instead of fixed-length n-grams [23]) are required to capture the specifics of the national language (i.e., a language that has unique syntactical features such as special letters, which are not present in other languages).

In text classification almost all words contain some information. Rudman [24] finds that more than 1,000 different style markers have been proposed. Different feature ranking methods can be applied to reduce the feature set, however, as Joachims [25] has demonstrated, even the features ranked lowest still contain considerable information and are relevant for classification. There is a significant amount of research still to be done in formulating and studying the language-specific features on all levels (syntactical, semantic, prosodic, etc.) such as the frequency of language-specific letters, frequency of n-grams with language-specific letters, forbidden n-grams, etc. [26]. Hereinafter we analyze the specific features of Lithuanian language texts.

The Lithuanian language is spoken by approximately 3.2 million people and is subject to numerous linguistic studies. It belongs to the Baltic group of the Indo-European family of languages. Lithuanian is considered an archaic language because it has preserved a lot of features otherwise found only in the ancient languages, such as Greek, Latin, and Sanskrit but which have disappeared in other modern languages. Such features are the preservation of Indo-European vowels and consonants, richness of inflection, preservation of old endings, and wide use of participial forms. The Lithuanian alphabet consists of the Latin alphabet letters (excluding Q, q, W, w, X, x) with eighteen extra letters with diacritics (nine capital and nine small). The Lithuanian language has thirty two letters, of which twelve are vowels (a, ą, e, ę, ė, i, į, y, o, u, ų, ū), six are semivowels (v, j, l, m, n, r) and 14 are consonants (b, c, č, d, f, g, h, k, p, s, š, t, z, ž). However, in electronic discourse, the letters with diacritics are very often replaced with matching Latin letters (e.g.: ą → a, č → c, ę → e, ž → z, etc.) or pairs of letters expressing the same sounds as in English (e.g.: č [tʃ] → ch, š [ʃ] → sh, etc.). There are nine diphthongs: ai, au, ei, eu, oi, ou, ui, ie, and uo. The principal feature of the Lithuanian language is the fact that the language has very many forms. Nearly all the inflectional parts of the language have 24–28 forms. E.g., the English word "two" has five forms in the nominative case alone, while there are thirty forms of the Lithuanian word "du" (= two) alone [27]. The Lithuanian language is particularly characterized by unusual richness in suffixes: there are 615 nominal suffixes, while in modern English there are only 113 nominal suffixes. On the other hand, the number of prefixes is not large: in Lithuanian there are only thirty six prefixes [28], while modern English has fifty seven prefixes. Lithuanian is highly inflective, ambiguous (47 per cent of words are ambiguous), has rich vocabulary (0.5 million headwords) and has complex word derivation system (e.g., seventy eight suffixes for diminutives) [29]. Verbs have 3 conjugations, and are inflected by four tenses, three persons, two numbers, and three moods. Non-conjugative forms of verbs retain the same root, but have different suffixes and endings in different inflection forms. There is a significant difference between frequency of unigrams in Lithuanian and in other (English, Polish, Serbian) languages (see Table I). Previous experiments in authorship identification using Lithuanian texts have demonstrated that content-features are more useful compared with function words or POS tags [29], while best results were obtained with word-level character tetra-grams and a set of lexical, morphological, and character features [30].

The promising directions of research are the use of unique character combinations in national languages, e.g., "eux" in French, "ery" in English, and "lj" in Serbo-Croatian [31], the use of language-specific function words such as "ale", "i", "nie", "to", "w", "z", "że", "za", "na" in Polish [32] and "neden-why", "ayrıca-furthermore", "belki-maybe", "daima-always" in Turkish [33], and the use of non-standard words such as abbreviations, acronyms [34].

Function words are words which serve to express grammatical relationships with other words within a sentence or to specify the attitude or mood of the speaker but do not have a specific lexical meaning. They can signify structural relationships between different words in a sentence. Function words might be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles. The frequency statistics for several languages is presented in Table II. Further we have analyzed and performed experiments with two subsets of textual features: one subset contains sets of language-independent stylometric features, which have been commonly used for authorship analysis of English texts as follows: number of words, number of lines, ratio of uppercase letters, frequency of numbers, frequency of white characters,

TABLE I
FREQUENCY STATISTICS (PERCENTAGE) (BASED ON DATA FROM HTTP://WWW.CRYPTOGRAM.ORG AND HTTP://MOFOBURRELL.LIVEJOURNAL.COM) OF TOP 5 UNIGRAMS IN 4 LANGUAGES

Order	English	Lithuanian	Polish	Serbian
1	e / 12.70	i / 15.25	e / 9.17	а / 10.99
2	t / 9.06	a / 10.43	o / 8.99	и / 8.43
3	a / 8.17	s / 9.34	i / 6.79	о / 8.15
4	o / 7.51	t / 6.75	a / 6.76	е / 8.03
5	i / 6.97	e / 5.55	y / 6.04	н / 4.64

TABLE II
FREQUENCY STATISTICS (PERCENTAGE) (BASED ON DATA FROM HTTPS://EN.WIKTIONARY.ORG AND HTTP://WWW.LEXITERIA.COM/) OF TOP 5 FUNCTION WORDS IN 4 LANGUAGES

Order	English	Lithuanian	Polish	Serbian
1	the / 4.90	ir - and / 3.32	w / 6.34	је - is / 4.23
2	be / 2.79	kad - that / 0.90	i - and / 2.56	у - near / 3.43
3	and / 2.39	į - to / 0.85	na - on / 2.03	и - and / 3.27
4	of / 2.30	iš - from / 0.67	z - from / 2.00	це - me /1.64
5	a / 2.25	su - with / 0.60	ię - themselves / 1.47	на - at /1.45

TABLE III
LEXICAL AND MORPHOSYNTACTIC FEATURES OF LITHUANIAN LANGUAGE

No.	Feature types	Examples of features and description
1	Function words	Frequency of each Lithuanian function word. Examples: “ant” (= on), “apie” (= about), “ar” (= whether), “arba” (= or), “aš” (= I), “be” (= without)
2	All function words	Cumulative frequency of all Lithuanian function words
3	Stop words	Frequency of each Lithuanian stop word. Examples: “į” (= into), “šalia” (= near), “šįjį” (= this, *masc.*), “šiąją” (= this, *fem.*)
4	All stop words	Cumulative frequency of all Lithuanian stop words
5	Word endings	Frequency of each Lithuanian language specific word ending. Examples: “a”, “ai”, “ajam”, “ame”, “ams”, “ant”
6	Uncommon bigrams	Frequency of each bigram uncommon to Lithuanian language. Examples: “qu”, “sh”, “zh”, “ch”, “ux”, “xu”
7	All uncommon bigrams	Cumulative frequency of all uncommon bigrams
8	Prefix “ne”	Frequency of words with prefix “ne” (= not)
9	Letters	Frequency of each Lithuanian language specific letter. Examples: “ą”, “č”, “ę”, “ė”, “į”, “š”
10	All letters	Cumulative frequency of all Lithuanian language specific letters
11	Abbreviations	Frequency of each Lithuanian language specific abbreviation. Examples: “gyd.” (= medical doctor), “kun.” (= priest), “tūkst.” (= thousand), “vyr.”(= senior)
12	All abbreviations	Cumulative frequency of all Lithuanian language specific abbreviation
13	Similes	Frequency of each Lithuanian simile. Examples: “pavyzdžiui” (= for example), “kaip” (= like), “tarkim” (= say)
14	All similes	Cumulative frequency of all Lithuanian similes.

frequency of letters, ratio of short (less than four letters) words, mean word length, number of sentences, mean sentence length, ratio of unique words, frequency of the most frequent word, ratio of delimiters, number of paragraphs, mean line length, frequency of endings, frequency of bigrams, ratio of unique bigrams, ratio of abbreviations, ratio of similes. Another one contains lexical features that are specific to Lithuanian language texts. The types of features used to calculate lexical features are summarized in Table III.

IV. Method

We perform authorship attribution using one-class classification. The one-class classification problem is the problem of distinguishing one class of objects from all others, given training data only for the target class. It has been introduced by [35] to handle training using only positive class information. As opposed to binary classification problems, here a boundary in the space of the objects of interest has to be inferred only from samples of positive class. One-class classification is often used for outlier or novelty detection because it attempts to differentiate between data that appears normal and data that appears abnormal with respect to training data.

We have selected Support Vector Machine (SVM) [36] based on comparative research indicating that SVM classifiers perform best on a variety of text classification experiments in the text analysis domain [25]. The One-Class SVM separates all the data points from the origin (in feature space V) and maximizes the distance from this hyper-plane to the origin. This results in a binary function which captures regions in the input space where the probability density of the data lives. Thus the function returns $+1$ in a region capturing the training data points and -1 elsewhere.

The classifier constructs a decision function F using the following *Decision function construction algorithm*:

```
ALGORITHM: constructDecisionFunction
INPUT: feature space V, training text set T^K
OUTPUT: decision function F
BEGIN
   FOREACH feature v IN feature space V of T^K
      IF value of feature v
         is predicted to arise from the
      distribution which generated
         the training samples of T^K
      THEN
         LET F(v) = 1
      ELSE
         LET F(v) = -1
      END IF
   END FOREACH
   RETURN F
END
```

For classification, we use the Sequential Minimal Optimization (SMO) algorithm [37] based implementation of one-class SVM classifier from DLIB C++ Library [38]. The classifier uses Radial Basis Function (RBF) kernel with default parameter values. The decision function F is used to create a ranked list of authors for each unknown text t as follows (see the following *Suspected author list construction algorithm*). Note that Heaviside function H is used to calculate the number of positive values of the decision function F.

```
ALGORITHM: createSuspectedAuthorList
INPUT: decision function F, testing text set T^L,
       list of authors A, number of suspects L
OUTPUT: ranked list of authors RL
BEGIN
   FOREACH a IN A
      FOREACH feature v IN feature space V of T^L
         LET LIST(a) = sum(H(F(v)))
            % H is the Heaviside function
      END FOREACH
   END FOREACH
   LET RL = sort(LIST)
      % return a ranked list of authors RL
   RETURN RL(1 : L)
END
```

V. Evaluation

We treat the author attribution system as the recommender system that using training data outputs a ranking order for the authors based on their predicted authorship relevance values. This approach known as Learning-to-rank is widely used in commercial search engines and recommender systems [39]. The evaluation of the authorship attribution system is not a trivial task. Commonly, such systems are evaluated using hard classification accuracy metrics such as precision and recall, which are often combined into a single measure such as F-score. These are set-based measures: authors in the ranking list are treated as unique and the ordering of results is ignored. We however claim that hard classification measures do not fit for the authorship attribution problem as they imply a strong oversimplification of reality. Therefore, we use soft classification measures based on the membership of a true author in a ranked list rather than direct match. When testing, the rank of the true author (which should be equal to equation (1) is to be compared with the predicted rank of the true author.

The simplest precision measure is to assume that we are only interested in the first suspected author and calculate the average probability of predicting the true author as the first author directly (Winner-Takes-All, WTA) [40] as follows:

$$WTA = \frac{1}{|T^K|} \sum_{T^K} H(rank(a_{true}) = 1) \qquad (1)$$

where $rank(a_{true})$ is the rank of the true author a_{true}, H is the Heaviside function, and T^K is the testing set of texts.

A more relaxed measure is to calculate if the list of the suspected authors contains the true author regardless of the position of the author within the suspected list. We call this metric List Precision (LP) and define in equation (2):

$$LP(L) = \frac{1}{|T^K|} \sum_{T^K} H(rank(a_{true}) \leq L) \qquad (2)$$

where L is the length of the list of suspected authors. Note that $LP(1) = WTA$.

Other measures used are based on evaluating ranked results, where importance is placed on returning a true author higher in the ranked list of suspected authors. These measures can be expressed using the gain-discount based model as a sum over authors in a ranked list as follows:

$$\pi \leftarrow \sum_{k=1}^{K} gain(k) \cdot discount(k) \tag{3}$$

where the $gain$ function represents the gain associated with the true author appearing at rank k, and the $discount$ function represents a discount associated with rank k, which is independent of the author, and K is the length of the suspected author list.

Equation (3) can be interpreted in terms of a simple model that simulates the work of a forensics expert: the expert starts with the first author and works his way down the list, eventually stopping [41]. The discount value indicates the probability that the expert continues his/her work at rank k, and the gain value represents the benefit (usability) to the expert of analyzing the author at rank k. Thus, the sum in equation (3) can be understood as expected total benefit experienced by the expert, with various gain values and discount formula corresponding to different assumptions about complexity of the expert's work.

Several different gain and discount functions have been proposed in the literature, which results in mean reciprocal rank [42], normalized discounted cumulative gain [43] and rank-biased precision [44] metrics. We describe the measures in brief below.

Mean reciprocal rank (MRR) is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of correctness. Here it is interpreted as the average of the inverse of the rank of the true author for a sample of test text dataset as in equation (4):

$$MRR = \frac{1}{|T^K|} \sum_{T^K} \frac{1}{rank(a_{true})} \tag{4}$$

Discounted cumulative gain (DCG) computes a value for the number of correctly recognized authors that includes a logarithmic discount function to progressively reduce the importance of authors placed further down the ranked list. This simulates the assumption that the experts will prefer the results which place the true author higher in the ranked list. The measure also makes the assumption that highly relevant authors are more useful than partially relevant authors, which in turn are more useful than non-relevant authors. DCG is defined in equation (5):

$$DCG = \frac{1}{|T^K|} \sum_{T^K} \frac{1}{log_2(rank(a_{true}) + 1)} \tag{5}$$

Rank-Biased Precision (RBP) (in equation (6)) assigns an effectiveness score to a ranking by computing a geometrically weighted sum of author relevance values, with the monotonically decreasing weights in the geometric distribution determined via a persistence p, $0 \leq p < 1$, where a smaller p value places greater emphasis on authors that appear early in the ranking, and a larger p spreads the weight further down the author ranking, but in both cases all authors in the ranking contribute to the final score.

$$RPB = \frac{1}{|T^K|} \sum_{T^K} p^{rank(a_{true})-1} \tag{6}$$

where p is an abstraction of the expert's searching persistence, expressed as a parameter between 0 and 1. Previous studies suggest that for web search a p value of 0.8 is an appropriate value, however, in practice, values as high as 0.95 are used [45].

VI. Results and discussion

The dataset was composed of Internet comments harvested from the Lithuanian news portal DELFI (http://www.delfi.lt) and covers the period of 8 months, from January, 2015 to August, 2015.

These comments were posted by anonymous users expressing their opinions about articles. Internet comments cover wide range of topics, are single units, do not necessary refer to each other; moreover, authors, hiding behind the anonymity curtain when expressing their opinions, have no reasons to pretend "better", therefore usually make no efforts to modify their writing style. But as the result of it, Internet comments are full of non-normative vocabulary words, include diminutives, hypocoristic words, and words with missing diacritics.

The authorship of the Internet comments was established based on an assumption that the identity of some author can be revealed, if his/her texts are written under the unique IP address and the unique pseudonym (taking both together as a single unit). Although some exceptions (when the same author is writing under several different IP addresses using different pseudonyms) may still occur, we anticipate they are too rare to make the significant influence on the overall authorship identification results.

Text fragments containing non-Lithuanian alphabet letters (except punctuation marks and digits) were eliminated; replies to comments and meta-information were discarded out as well leaving just plain texts. Besides, the texts shorter than thirty symbols (excluding white-space characters) were not included. Finally, all texts by the same author were concatenated, yielding the texts consisting of between 3,543 and 119,169 symbols. The composed corpus contains the texts of 200 authors. While the corpus is not very large, it is leger than datasets commonly used in the authorship forensics domain, e.g., [46] use only 300-word texts of three authors. The characteristics of the dataset (length of the longest, mean, and shortest text message in characters and words) are summarized in Table IV.

In our experiment, we have randomly selected 80 per cent of each author texts for training, while the remaining 20 per cent of texts together with texts of other authors were used for testing. The results of one-class classification were used for constructing a list of suspected authors. First, we ranked

TABLE IV
CHARACTERISTICS OF THE DATASET

Characteristic	Value
Number of authors	200
Number of texts	200
Length of shortest text, characters	3,543
Length of shortest text, words	504
Length of longest text, characters	119,169
Length of longest text, words	15,874
Average length of text, characters	15,866
Average length of text, words	2,267

authors based on author attribution using general features only. Next, we added the language-specific (Lithuanian, LT) features and to see if there was an improvement of the position of the true author in the rank of the suspected authors. For evaluation, the process was repeated for each of 200 authors.

Fig. 1 presents the results of experiments (in terms of List Precision) using only general text features as well as general and language-specific features. The results demonstrate a marked improvement in precision when Lithuanian language-specific lexical features have been added for classification. For comparison, a random baseline, which represents the probability that the true author will be assigned to the suspected author list, is also shown.

The evaluation of experimental results using the gain-discount model based rank evaluation metrics is given in Table V (mean values are given). To demonstrate the efficiency of the one-class classification approach for ranking we compare the values of metrics with the random guess baseline, which shows the lowest accuracy threshold which that be exceeded that the applied approach could be considered as effective and reasonable enough for author attribution tasks. An improvement of accuracy achieved using language-specific lexical features is given with 95 percent statistical confidence interval (mean $\pm$ 1.96 $\cdot$ standard deviation). The paired t-Student confirmed (at 0.05 level) that the differences between the obtained results were statistically different for all considered metrics ($p < 10^{-8}$).

Finally, we present the evaluation of language-specific subsets of lexical features by calculating the average improvement in the rank position of the true author (see Fig. 2).

The results obtained (see Fig. 2) show that the best improvement is achieved using language-specific function words (column 1), word endings (column 5) and stop words (column 3).

These results are consistent with the findings of authors using function words as a reliable base for textual comparison, which are not strongly affected by a text's topic or genre, or an author's conscious control while writing [47].

Word endings have been noted to contribute to the success of character n-grams in stylometric analysis [48], however, in Lithuanian language the word endings are typically longer than bigrams or trigrams commonly used as general features (we used bi-grams only), thus a separate feature type of Lithuanian word endings seems to be useful.

Stop words are a very strong indication of writing style that convey very little semantic meaning in a sentence but serve to add details to it. Stop words on the other hand are inevitable in the output of any author and hence a generalizable technique cannot but tap their properties. Moreover stop words are result of a subconscious process of constructing sentences and thus may serve as a writeprint of the authors [49].

The use of features based on language-specific letters of alphabet (column 9) yielded negative results due to the non-normative use of such letters in the electronic space, e.g., replacement with similar Latin alphabet letters without diacritics or with similarly sounding English bigrams.

VII. CONCLUSIONS

In this paper we have presented a one-class classification based method for open-class authorship attribution. The method produces the list of suspected authors rather than a single predicted author, therefore, reducing the problem from being on a large scale to smaller scale. The method could be used by the forensic linguistics expert community to help identify the list of suspects to be analyzed further using manual methods. The proposed method allows to reduce the number of suspected authors by fourfold (from 200 to 50) with a probability of 0.90 and eightfold (from 200 to 25) with a probability of 0.80 that the true author is listed as the suspected author.

We have discussed the metrics for evaluation of the result and suggested using rank correlation based metrics. We have proposed the List Precision metric to evaluate the usability of the derived suspected author list based on the length of the list. We also have identified language-specific lexical features.

The experimental results using the online Lithuanian language texts (dataset of online forum comments) classified using one-class SVM classifier show that Lithuanian function words together with Lithuanian word endings and stop words are the ones which contribute most towards the improvement of the classification results (0.13-0.17, based on different evaluation metrics).

The results were evaluated statistically using paired t-Student test showing that the improvement in the value of usability metrics was statistically significant.

In future research we are planning to increase the number of authors in the datasets; to analyze different domains (e.g.

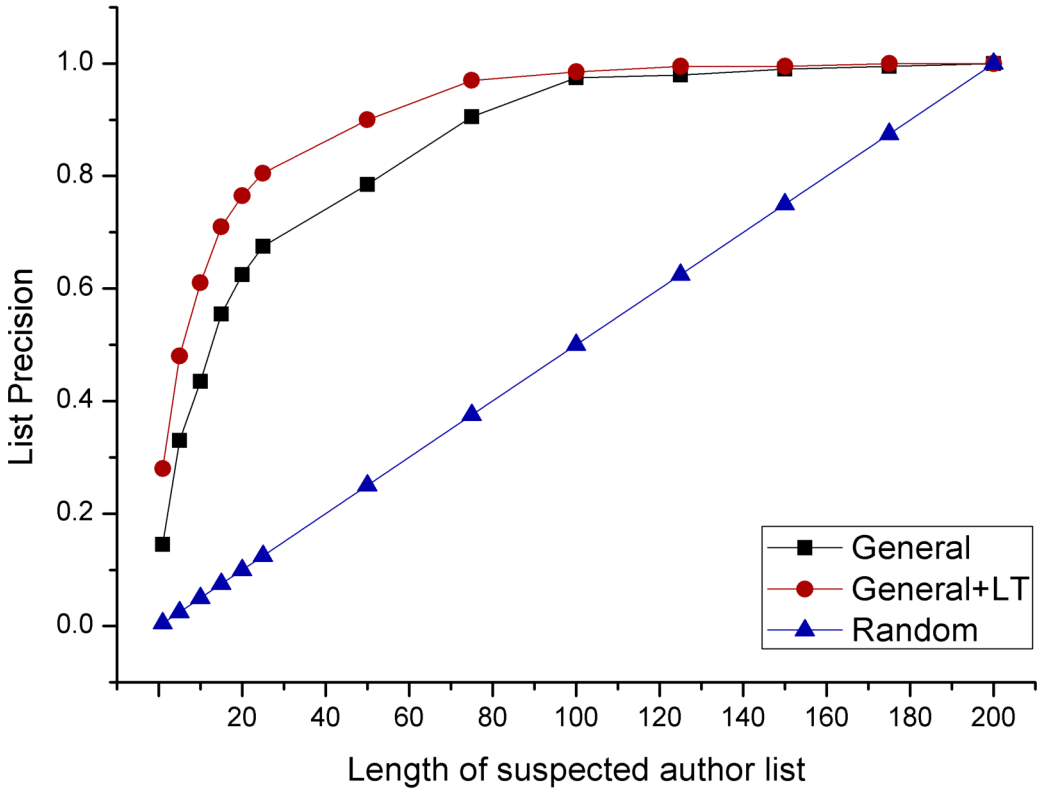

Fig. 1. Accuracy of general and language-specific features using List Precision measure

TABLE V
EVALUATION RESULTS USING RANK CORRELATION BASED METRICS

Metric	Random baseline	Without language-specific (LT) features	With language-specific (LT) features	Improvement in accuracy (with 95% stat. confidence limits)
WTA	0.005	0.145	0.280	0.135±0.054
LP (L=10)	0.055	0.435	0.610	0.175±0.068
MRR	0.025	0.237	0.391	0.154±0.042
DCG	0.169	0.385	0.513	0.131±0.034
RPB (p=0.8)	0.016	0.288	0.455	0.167±0.043

blogs, tweets, etc.) and language types as well as to focus on sentiment-related lexical features, and analyze novel semantic feature descriptors such as Holomorphic Chebyshev Projectors [50].

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of the project "Lithuanian Cybercrime Centre of Excellence for Training, Research & Education" (L3CE) project, Grant Agreement No. HOME/2013/ISEC/AG/INT/4000005176), financed by European Commission under the Programme EC DG Home Affairs ISEC (Prevention of and against crime 2007-2013).

REFERENCES

[1] Irons, A., and Lallie, H.S. 2014. Digital Forensics to Intelligent Forensics. *Future Internet*, 6, 584-96.

[2] Chaski C. E. 2012. Author Identification in the Forensic Setting. In L. Solan and P. Tiermsa (Eds.), *The Oxford Handbook of Forensic Linguistics*, Oxford University Press.

[3] Iqbal, F., Binsalleeh, H., Fung, B. C. M., and Debbabi, M. 2013. A unified data mining solution for authorship analysis in anonymous textual communications. *Inf. Sci., Vol.*, 231, pp. 98–112.

[4] Koppel, M., Schler, J., and Argamon, S. 2011. Authorship Attribution in the Wild. Language Resources and Evaluation, 45(1), pp. 83–94.

[5] Van Halteren, H. 2004. Linguistic profiling for authorship recognition and verification. *Proc. of 42nd Meeting on Association for Computational Linguistics*, ACL'2004, pp. 199–206.

[6] Brocardo, M. L., Traore, I., and Woungang, I. Authorship verification of e-mail and tweet messages applied for continuous authentication. *J. Comput. Syst. Sci.*, 81(8), pp. 1429–40.

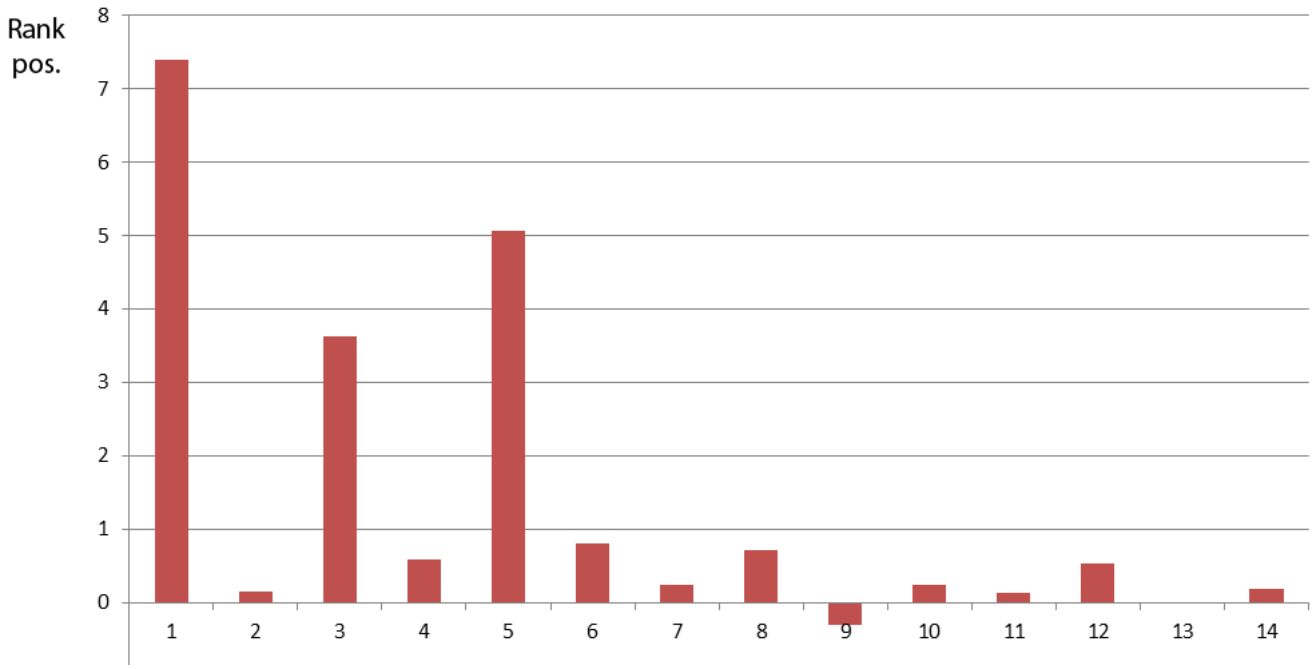

Fig. 2. Improvement in average rank position of a true author using Lithuanian specific features: 1-Frequency of function words, 2-Frequency of all function words, 3-Frequency of stop words, 4-Frequency of all stop words, 5-Frequency of words with specific endings, 6-Frequency of uncommon character bigrams, 7-Frequency of all uncommon character bigrams, 8-Frequency of prefix "ne", 9-Frequency of letters, 10-Frequency of all letters, 11-Frequency of abbreviations, 12-Frequency of all abbreviations, 13-Frequency of similes, 14-Frequency of all similes

[7] Neralla, S., Bhaskari, D.L., and Avadhani, P. S. 2014. A Stylometric Investigation Tool for Authorship Attribution in E-Mail Forensics. In *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol. II. Advances in Intelligent Systems and Computing*, Vol. 249, pp. 543-9.

[8] Alazab, M., Layton, R., Broadhurst, R., and Bouhours, B. 2013. Malicious Spam Emails Developments and Authorship Attribution. *Proc. of 4th Cybercrime and Trustworthy Computing Workshop (CTC '13)*, pp. 58–68.

[9] de Vel, O., Anderson, A., Corney, M., and Mohay, G. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4), pp. 55–64.

[10] Potthast, S., Stein, B., Barron-Cedeno, A., and Rosso, P. 2010. An Evaluation Framework for Plagiarism Detection. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pp. 997–1005. ACL.

[11] Amuchi, F., Al-Nemrat, A., Alazab, M., and Layton, R. 2012. Identifying Cyber Predators through Forensic Authorship Analysis of Chat Logs. *Third Cybercrime and Trustworthy Computing Workshop (CTC)*, pp. 28-37.

[12] Damasevicius, R., Valys, R., and Wozniak, M. 2016. Intelligent tagging of online texts using fuzzy logic. *IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 1-8. IEEE.

[13] Krilavičius, T., Medelis, Z., Kapočiūtė-Dzikienė, J., and Žalandauskas, T. 2012. News Media Analysis Using Focused Crawl and Natural Language Processing: Case of Lithuanian News Websites. *Proc. of Int. Conf. on Information and software technologies*, ICIST 2012, pp. 48–61.

[14] Steen, R. G. 2014. The Demographics of Deception: What Motivates Authors Who Engage in Misconduct? *Publications*, 2, 44-50.

[15] Ding, S.H.H, Fung, B.C.M., and Debbabi, M. 2015. A Visualizable Evidence-Driven Approach for Authorship Attribution. *ACM Trans. Inf. Syst. Secur.*, 17, 3, Article 12, 30.

[16] Koppel, M., and Schler, J. 2004. Authorship Verification As a One-class Classification Problem. *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*, 489-495.

[17] Veenman, C. J., and Li, Z. 2013. Authorship Verification with Compression Features. Working Notes for CLEF 2013 Conference. *CEUR Workshop Proceedings* 1179.

[18] Can, M. 2014. Authorship Attribution Using Principal Component Analysis and Competitive Neural Networks. *Math. Comput. Appl.*, 19, 21-36.

[19] Mikros, G., and Perifanos K. 2013. Authorship attribution in greek tweets using author's multilevel n-gram profiles, in: *AAAI Spring Symposium Series*.

[20] Sousa-Silva, R., Sarmento, L., Grant, T., Oliveira, E., and Maia, B. 2010. Comparing sentence-level features for authorship analysis in Portuguese. *Proc. of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR'10)*, pp. 51-54.

[21] Reicher, T., Krišto, I., Belša, I., Šilic, A. 2010. Automatic authorship attribution for texts in Croatian language using combinations of features. *Proc. of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part II (KES'10)*, pp. 21-30.

[22] Graovac, J. 2012. Serbian Text Categorization Using Byte Level n-Grams. *Local Proceedings of the Fifth Balkan Conference in Informatics, BCI'12*, pp. 93–96.

[23] Tomović, A., and Janičić, P. 2007. A Variant of N-Gram Based Language Classification. *Artificial Intelligence and Human-Oriented Computing, 10th Congress of the Italian Association for Artificial Intelligence, AI*IA 2007*, pp. 410–421.

[24] Rudman, J. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, pp. 351–365.

[25] Joachims, T. 2002. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, Norwell, MA, USA.

[26] Venčkauskas, A., Damaševičius, R., Marcinkevičius, R., and Karpavičius, A. 2015. Problems of authorship identification of the national language electronic discourse. In: *Proc. of the 21st Int. Conference on Information and software technologies, ICIST 2015*, pp. 415–432.

[27] Šveikauskienė, D. 2005. Graph Representation of the Syntactic Structure of the Lithuanian Sentence. *INFORMATICA*, Vol. 16, No. 3, pp. 407–418.

[28] Klimas, A. 1974. Studies on Word Formation in Lithuanian. *Lituanus Lithuanian Quarterly Journal of Arts and Sciences*, 20(3).

[29] Kapočiūtė-Dzikienė, J., Utka, A., and Šarkutė, L. 2014. Feature Exploration for Authorship Attribution of Lithuanian Parliamentary Speeches.

Proc. of 17th International Conference on Text, Speech and Dialogue, TSD 2014, pp. 93–100.

[30] Kapočiūtė-Dzikienė, J., Utka, A., and Šarkutė, L. 2015. Authorship Attribution of Internet Comments with Thousand Candidate Authors. *Proc. of the 21st Int. Conference on Information and software technologies, ICIST 2015*, pp. 433–48.

[31] Zečević, A., and Stanković, S. V. 2013. Language Identification: The Case of Serbian. *Proceedings of Natural Language Processing for Serbian - Resources and Application*.

[32] Stahczyk, U., and Cyran, K. A. 2007. Machine learning approach to authorship attribution of literary texts. *Journal of Applied Mathematics*, 7(4):151–8.

[33] Türkoğlu, F., and Diri, B. 2007. Fatih Amasyali, M. Author attribution of Turkish texts by feature mining. *Proc. of the 3rd international conference on Advanced intelligent computing theories and applications (ICIC'07)*, pp. 1086–93.

[34] Beliga, S., and Martincic-Ipsic, S. 2014. Non-standard words as features for text categorization. *37th Int. Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014*, pp. 1165–9.

[35] Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. 1999. Support Vector Method for Novelty Detection. *Advances in Neural Information Processing Systems 12, NIPS 1999*, pp. 582–8.

[36] Vapnik, V. N. 1995. *The nature of statistical learning theory*. Springer-Verlag.

[37] Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 27.

[38] King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10, 1755–8.

[39] Li, P., Burges, C., and Wu, Q. 2007. McRank: Learning to rank using classification and gradient boosting. *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, NIPS*, pp. 897–904.

[40] Chapelle, O., Le, Q., and Smola, A. 2007. Large margin optimization of ranking measures. In *NIPS Workshop on Machine Learning for Web Search*.

[41] Smucker, M. D., and Clarke, C. L. A. 2012. Time-based calibration of effectiveness measures. *Proc. of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*, pp. 95–104.

[42] Craswell, N. 2009. Mean Reciprocal Rank. *Encyclopedia of Database Systems*, Vol. 1703.

[43] Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4), pp. 422–46.

[44] Moffat, A. and Zobel, J. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM TOIS*, 27(1), pp. 1–27.

[45] Zhang, Y., Park, L. A. F., and Moffat, A. Parameter sensitivity in rank-biased precision. *Proc. of the 13th Australasian Document Computing Symposium (ADCS)*, pp. 61–68.

[46] Nini, A., and Grant, T. 2013. Bridging the gap between stylistic and cognitive approaches to authorship analysis using Systemic Functional Linguistics and multidimensional analysis. *International Journal of Speech Language and the Law*, 20(2), pp. 173-202.

[47] Kestemont M. 2014. Function Words in Authorship Attribution From Black Magic to Theory? *Proc. of the 3rd Workshop on Computational Linguistics for Literature (CLfL) at 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014*, pp. 59–66.

[48] Forstall, C., and Scheirer, W. 2010. Features from Frequency: Authorship and Stylistic Analysis Using Repetitive Sound. *Proc. of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(2).

[49] Arun, R., Suresh, V., and Veni Madhavan, C. E. 2009. Stopword Graphs and Authorship Attribution in Text Corpora. *Proc. of the 2009 IEEE International Conference on Semantic Computing (ICSC '09)*, pp. 192–6.

[50] Napoli, C., Tramontana, E., Lo Sciuto, G., Woźniak, M., Damaševičius, R., and Borowik, G. 2015. Authorship Semantical Identification using Holomorphic Chebyshev Projectors. In: Asia-Pacific Conference on Computer Aided System Engineering (APCASE), pp. 232-237. IEEE.