Finding Participants in a Chat: Authorship Attribution for Conversational Documents

Giacomo Inches, Morgan Harvey and Fabio Crestani
Faculty of Informatics
University of Lugano (USI)
Lugano, Switzerland
{giacomo.inches, morgan.harvey, fabio.crestani}@usi.ch

***Abstract*—In this work we study the problem of Authorship Attribution for a novel set of documents, namely online chats. Although the problem of Authorship Attribution has been extensively investigated for different document types, from books to letters and from emails to blog posts, to the best of our knowledge this is the first study of Authorship Attribution for conversational documents (IRC chat logs) using statistical models. We experimentally demonstrate the unsuitability of the classical statistical models for conversational documents and propose a novel approach which is able to achieve a high accuracy rate (up to 95%) for hundreds of authors.**

I. Introduction and motivations

Studies on the problem of Authorship Identification and its variants, from Authorship Attribution to Profiling, are not new and have been conducted quite extensively in recent years. Despite the abundance of previous work on Authorship Identification (the first of which is dated 1887 [1]), the problem of Authorship Identification is still of a great interest due to the availability of novel datasets and new techniques. While in the past researchers concentrated their efforts on collections of formal edited documents, like letters or newspaper articles [2], in recent years attention has moved to novel kinds of documents, such as emails or online conversations. We consider these documents novel because they have not been analysed in the past and present characteristics that makes them more difficult to analyse. One of these characteristic is the informal nature of the text contained within the documents, due to the fact that they are produced directly by some authors in an informal context (e.g. IRC chats or email) or not validated through a review process (e.g. blogs). Furthermore, conversational documents are by nature short, unstructured and full of spelling mistakes [3].

The yearly PAN[1] laboratory competition is an example of such ongoing research in the field of Authorship Identification [4], [5], with some specific and innovative tasks like the "predator identification in online conversations" subtask (PAN 2012) or the "users profiling in social media" task (PAN 2013). This data is interesting for researchers because it presents different challenges when analysing it, from preprocessing (how to expand short documents? how to deal with dozens of spelling mistakes?) to analysis (how to profile users?) but also for companies, in particular those providing services for facilitating user conversations and collaboration. For example, providers of the IRC service or Social Networks might want to be able to identify particular users based on their behaviour (or misbehaviour) or be able to profile them to better target advertisements.

For this reason in our work we focus on a set of such novel documents, in particular ones of a conversational nature. To the best of our knowledge this is the first work that studies the problem of Authorship Attribution in IRC chat logs using statistical models. IRC chat logs are not only good representatives of the class of novel documents due to their informal, unstructured and conversational nature, but also contain hundreds of different authors, that makes the problem of identifying them more difficult. The main contributions of this work are the following:

- we demonstrate the unsuitability of standard approaches for Authorship Attribution when applied to novel conversational documents, like IRC chats logs;
- we identify the best existing statistical models for Authorship Attribution in this setting;
- finally, we extend the best statistical models, making use of the conversational nature of our documents, to improve the accuracy in the classification of the authors up to 95%.

The paper is organised as follows: in Section II we present the related work and an introduction to the problem of Authorship Identification in all its variants, in Section III we illustrate the traditional models for Authorship Attribution and describe our proposed model for conversational documents. In Section IV we describe the publicly available collection used for our experiments, the experimental settings and the results obtained with the different models. We conclude in Section V by highlighting possible extensions of the proposed methods and future research directions emerging from this study.

II. Related work

Good introductory works on the topic of Authorship Identification are a book by Juola [2] and an article by Stamatatos [6], where the two authors highlight the main techniques and applications of Authorship Identification. These techniques generally apply one of two different approaches to the classification problem, namely generative (e.g. Bayesian) models and discriminative (e.g. Support Vector Machine) models.

[1] Evaluation lab on uncovering plagiarism, authorship, and social software misuse http://pan.webis.de

In combination with these classification approaches, different features can be used to characterise the authors: from lexical and character features, to syntactic and semantic ones. Two classes of Authorship Identification problem are traditionally identified with the expression Authorship Attribution: “closed class” and “open class”. In the “closed class”, given a text, one should attribute it to a single author from a predefined group, where the training and testing sets are the same. In the “open class” problem, however, the set of possible authors may not be limited to a predefined subset but may involve other authors from outside the predefined set. The third class of problem is mostly referred to Profiling (or Stylometry [7]) and focuses on identifying properties of the authors of a given text, such as age, sex, dialect, etc. [8], [9].

As mentioned in Section I, online conversations and social media are two means of communication for which little research has been done in the context of Authorship Identification. Apart from the specific task of predator identification [10], there is little research explicitly addressing the problem of Authorship Attribution [11] or Profiling (Stylometry) in online conversations [12], [13], with many studies concentrating on chat disentanglement or segmentation [14]–[16]. The same situation can be observed for social media, where only few publications exist on the topic of Authorship Identification. In this context, some studies have been conducted covering newsgroups [17], blogs [18], microblogs [19] and “real” social media like Netlog[2] [20].

It is also noticeable that online conversations differ significantly from social media like blogs, newsgroups or discussion fora for example in the length of the messages and in their style, as was demonstrated in previous studies [3], [21]. For this reason, the problem of Authorship Identification for online conversations should be approached in a different way compared to social media. In social media, in fact, discriminative approaches like SVM have been successfully employed [17], [18], [20] and have even been shown to be partially successful for the more specific problem of conversational content [10], but at some cost. First, the contribution of the different features toward the final result is not clear in case of SVM and there are applications where it is of primary importance that the feature which influenced the decision of the classifier can be identified [10], [22]. Moreover, SVM require a phase of training based on dedicated data and another phase of classification based on the model derived from the training sample. This limits the possibility of adapting existing models to new sets of data (e.g. new authors) without re-training or, even worse, it might be impossible to employ them at all due to a lack of training examples. Generative models, instead, seem to be more flexible and performant in the case of conversational content [12], [13].

Recent work by Savoy [23] explores, in detail, the most common statistical methods for Authorship Attribution, showing their suitability in comparison to other standard generative models (e.g. Naïve Bayes). Since these methods are as powerful and flexible as the generative models, but benefit further by allowing control over the contribution of each feature (term) in each document (author), we decide to make use of two of these approaches in our work. To conclude, out intention is to focus on statistical approaches that work at term level, instead of a character level [11], and that take hundreds of authors into consideration at the same time, rather than a small set of maximum 50 [11], [23]. Furthermore, we do not restrict the number of authors to those under investigation, like in the “closed” case, but we also allow non-relevant authors to be present as in the “open” case, which is clearly more difficult.

III. Statistical models for authorship attributions

In this section we present two traditional statistical models used for Authorship Attribution (Section III-A) and their extension, considering the conversational nature of the documents in our datasets (Section III-B). In Authorship Attribution the problem we are given is to decide which of the author profiles (A_j) is most similar to a given unknown profile, which we can consider an input “query” (Q).

A. Traditional Approaches

Traditional statistical models make use of terms and term frequencies in the texts to determine the similarity between them. In the following section we present two of the most used and effective statistical models [23]: Chi-squared (χ^2) distance and Kullback-Leibler divergence. We will use these models primarily to compute the similarity between two author profiles but also as tools to derive an author-specific set of terms to be used as his/her profile (Section IV).

1) χ^2 distance:

$$\chi^2(\mathrm{Q}, \mathrm{A}_j) = \sum_{i=1}^{m} \frac{\big(\mathrm{q}(t_i) - \mathrm{a}_j(t_i)\big)^2}{\mathrm{a}_j(t_i)} \qquad (1)$$

The χ^2 distance as presented in [23], [24] is displayed in Equation 1, where t_i is the relative term frequency of term i in the “query” document q and in each of the reference documents a_j. The origin of χ^2 is in the field of probability theory and statistics, where it is typically used to measure the difference between observed data and expected data. The greater the difference, the more the observed data diverges with respect to the expected data, thus one can conclude that the two sets of data are not related. We are using the χ^2 distance with the same intuition in this study, using one user profile as a “query” and measuring its distance to each candidate user profile. The less distant the two profiles, the more probable it is that the same author generated them. Each author profile is composed of terms that form a distribution, which can vary from author to author and from setting to setting, depending on the assumption we are making to build each profile. For example, the total number of terms $i = 1 \dots m$ depends on the assumption of the minimum document frequency for each term. In the original formulation [23], [24] this was tested at different levels (2, 5, 10) which we will

[2] www.netlog.com

discuss in Section IV-B1. Having computed the χ^2 distance between a "query" and all the user profiles, we minimize it to find the most likely profile, from the assumption that the distance between the profiles is minimized when they are most similar (equal or generated by the same author).

2) Kullback-Leibler Divergence:

$$\mathrm{KLD}(\mathrm{Q}||\mathrm{A}_j) = \sum_{i=1}^{m} p_q(t_i) \cdot \log_2 \left[\frac{p_q(t_i)}{p_j(t_i)}\right] \tag{2}$$

The Kullback-Leibler Divergence (KLD) (or relative entropy) is an asymmetric measure of disagreement[3] between two probabilistic distribution, which id derived directly from the concept of entropy [25], [26]. Analogously to χ^2, if two distributions were generated by the same process, or if two user profiles were generated by the same user, their dissimilarity, thus their KLD distance, will be minimal. For this reason we compute the KLD between the "query" document (the unknown user profile) and all the other profiles and then minimise this distance to find the closest profile, thus the user most associated with the query.

In previous work it has been demonstrated that KLD is an effective indicator of the similarity between two texts [2] and that it can be used successfully to address the Authorship Attribution problems [27]–[30]. In Equation 2 we indicate with $p_q(t_i)$ the probability of a particular term t_i in the "query" document q, while $p_j(t_i)$ identifies the probability of the same term t_i in a reference document j.

To estimate the probability of a term in a document, we first adopt the Maximum Likelihood Estimation (MLE), under which the probability of a term in a particular document is equal to the frequency of the term in the document (tf_i) divided by the total number of tokens in that document (n). Beside MLE, we also adopt a smoothing technique [26] based on Lindstone's law, as suggested in [23]. This allows null probabilities to be discarded due to the absence of the terms (when working on limited words sets, like in our case) and prevents probabilities from going to infinity due to a null denominator. We decided not to investigate other smoothing techniques, such as the Laplace smoothing or Dirichlet smoothing and leave them to future study.

$$p(t_i) = \frac{tf_i + \lambda}{n + \lambda \cdot |V|} \tag{3}$$

In Equation 3 we represent the full formula of the MLE with the smoothing parameter λ, adjustable as desired, and $|V|$ the vocabulary size of the entire corpus. It should be noted that for $\lambda = 0$, Equation 3 represents the formula for computing MLE alone and for $\lambda = 1$, Equation 3 represents the formula of Laplace smoothing.

To conclude, we give an overview of some established procedures for doing Authorship Attribution using χ^2 and KLD methods. The first step is building users profiles. This is usually achieved by concatenating all the texts written by the same author into a single document, which become the author profile. The author profile is then compared to all the possible queries, typically other unlabelled user profiles. The comparison is done using χ^2 and KLD, minimising the distances to obtain the best matching author profile, thereby finding an author for the unlabelled profile. We must now define the "support" (i.e. list of features or terms) to be used when comparing the user profiles. One strategy that was demonstrated to work well for determining such support is the use of stopwords or most common words across all user profiles [2], [23]. This allows the style of writing of each author to be identified, at least for written documents, and to treat each profile as a distribution over predefined terms. In our experiments in Section IV-B1 we test the suitability of this traditional approach on a novel set of documents, which are non-standard for Authorship Attribution, namely online conversations.

Due to the fact that online conversations have properties not found in standard documents, we first test the suitability of a simple approach considering as support all the terms in a user profile (results in Section IV-B2) we then propose a novel method based on the mutual influence of authors in a conversation for generating better support. We present this method in the following Section III-B and the associated experimental results in Section IV-B3.

B. Proposed Approach: Mutual Influence of Author Vocabulary in a Conversation

We propose a method for selecting the most discriminant words for each author based on his/her conversations and later use these words as support when comparing user profiles. This approach takes into consideration the nature of our dataset (i.e. composed of chat messages, an example of conversational documents) which presents some particular characteristics. An interesting list of properties of conversational documents can be found in [31], but there are two in particular that motivate our method:

- the property that a user message has an impact on all the future messages in a conversation and
- the fact that users need ways of emulating the non-verbal expressions that can be found in a regular in-presence conversation, thus creating a own novel language style, no longer related to the common language of all the users.

In formulating our approach we focus exclusively on the first property, while we make use of the second in the experimental setting in Section IV-B3.

Our approach can be divided into two parts. In the first part we analyse all the conversations in our dataset to find the group of users participating in the same conversation. We ignore uninformative conversations with a single interlocutor and those with too many users (more than 140), retaining 90% of the available data.

For each user we consider the list of the other users he/she talked to and generate a new list of profiles corresponding to each couple of users with their own respective joint term

[3] It is not a distance in the strict sense because it is not symmetric and does not respect the triangular inequality [25].

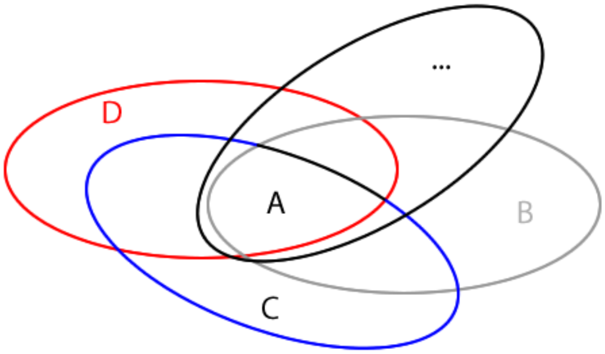

Fig. 1. Specific vocabulary for author A vs. {AB, AC, AD, ...}

usage. This procedure allows the vocabulary of one user to be influenced by the vocabulary of the other users he/she is talking with. For example, if user A is talking with B and C, we will generate the profiles AB (all the messages of A in that conversation are merged with the messages of B in that conversation) and AC, but also BC since both B and C also participated in the same conversation. In this formulation AB and BA are equal and mutually exchangeable, so we do not distinguish between A sending a message to B or B sending a message to A. In other words we are not considering the temporal aspect of the conversation, that we leave to future studies. Once we have generated all the couples for one author, we generate an extended profile for that author. As in the previous example, if for user A we have couples AB, CA, DA etc, then we merge them together and call it A*. The intuition behind this procedure is illustrated in Figure 1: all the couples have A in common and our goal is to identify the intersection: i.e. the specific vocabulary of A among all his/her conversations.

To identify the specific vocabulary of a user, we must now compute the dissimilarity between the combined profile and all his/her interlocutors' user profiles. For example, if A*, B*, C* and D* are the user profiles and we want to know the dissimilarity of a particular user A* and his/her interlocutors B*, C*, D*, we first generate the set Γ based on his/her interlocutors' profiles. We obtain Γ by merging his/her interlocutors' vocabularies and frequencies and have as a result: $\Gamma_{A*} = \{$B*, C*, D*$\}$. At this point we can compute the divergence Δ between each user profile and its connected Γ: e.g. $\Delta(\text{A*}||\Gamma_{A*})$ with $\Gamma_{A*} = \{$B*, C*, D*$\}$, etc. The divergence is computed at term level, using all terms in each profile. We use KLD (Section III-A2) as the measure of divergence Δ.

$$\text{KLD}(\text{A*}||\Gamma_{A*}) = \forall t \in \text{A*}, p_{A*}(t) \cdot \log_2 \left[\frac{p_{A*}(t)}{p_{\Gamma_{A*}}(t)} \right] \quad (4)$$

In Equation 4 we provide an example of how to compute the vocabulary specific to user A. For all the terms t belonging to the user profile A* we compute the KLD, where $p_{A*}(t)$ is the probability of term t in the user profile A* and $p_{\Gamma_{A*}}(t)$ is the probability of the term t in the collection Γ_{A*}. The probability is estimated through MLE with Lindstone smoothing as in Equation 3, with $\lambda = 0.1$ and $|V| = |\Gamma|$. At this point we have for each author a list of words, from the most to the least discriminant word. This list encapsulates also the influence the vocabulary of other persons chatting with the user has on his/her vocabulary. In Section IV-B3 we present the experimental results emerging from this method.

IV. Experimental evaluation

In this section we present the settings employed in our experimental framework, describing the datasets used and describing in detail the different choices, in particular regarding the process of term selection and author characterisation, which are central to our approach.

A. Dataset

Since we are interested in conversational documents, we use a subset of one of the most recent and complete collections of online conversation [10]. This collection incorporates 4 different dataset of IRC logs ("perverted justice", "krjin", "irclogs" and "omegle") and was originally designed to solve the problems of i) finding users that manifest unacceptable or illegal behaviour (such as a sexual predator) among a set of conversations and ii) to identify the lines of the conversations where this behavior manifested. In this work we are not interested in exploring these kind of problems (or simply profiling a class of users) but, instead, we want to characterise each author in any conversation. Moreover, for one dataset ("omegle") in the collection, every author has just one document produced and of a limited length, while for another one ("perverted justice") the conversations were between only two users. For these reasons, we decided to only use the two remaining datasets ("krjin" and "irclog"). We consider the two dataset separately and perform individual experiments on each of them. Although our study does not focus directly on topicality, we noticed that the topicality of the two datasets was to some extent homogeneous. Documents from "krjin", in fact, are centered on topics related to HTML 5 (e.g. html5, css, micro formats, accessibility, ...) while the ones from "irclog" are somewhat more diverse, ranging from java, gentoo and macosx to php, oracle and samba (see Appendix for the complete list of topics). Despite this homogeneity, we noticed, via manual inspection, that in many cases users engaged in conversations that diverged from the expected topic, discussing, for example, families, general interests and sometimes even anger.

In Table I we report the statistics for each of the two collections employed, where it is evident that they are similar in terms of average profile length, despite the fact the one contains manifestly more users and documents than the other. For each dataset we also extracted a subset of 20 users, that we used in our experiments to simulate the traditional case of Authorship Attribution where there is a limited number of users. Since traditional documents (like newspaper, letters, etc) are clearly longer than conversational ones, we selected our 20

TABLE I
DATASET STATISTICS

Dataset	# Docs.	# Users	Avg. Profile Length # Tokens	 # Singleton
irclogs	93327	4646	58.07	39.05
krijn	78605	19046	60.80	39.90
20 irclogs	13282	20	139.02	81.68
20 krijn	3247	20	136.74	78.01

TABLE II
RESULTS USING ONLY "STOPWORDS" AS USER PROFILE

Number of User Train Set	 Test Set	Collection	Model	TF	NIDF	Indri
20	20	irc logs	kld	0.80	0.80	0.73
			chi2	0.62	0.62	0.58
		krjin	kld	0.92	0.92	0.85
			chi2	0.67	0.67	0.75
400	400	irc logs	kld	0.14	0.14	0.16
			chi2	0.08	0.08	0.10
1501	1501	krjin	kld	0.09	0.09	0.08
			chi2	0.05	0.05	0.04
1868	400	irc logs	kld	0.12	0.12	0.13
			chi2	0.07	0.07	0.08
7580	1502	krjin	kld	0.08	0.08	0.07
			chi2	0.04	0.04	0.04

users in a random fashion whilst trying to preserve an average profile length comparable to standard documents.

B. Results

To identify users among conversations we adopt the scoring and classifying methods presented in Section III. In all the experiments reported we use a 3-fold-cross validation strategy: we split the whole dataset into 3 parts and we iteratively used 1/3 of the data as the testing set with "queries" (authors to be identified) and the remaining as reference or training set (authors to compare the query with and to compute the proposed method on). While we acknowledge that this is a somewhat unorthodox number of folds, it was chosen in order to ensure a sufficiently large "query" set. We perform experiments with the subset of 20 users and, unless otherwise stated, using half of the documents in each collection to keep the computational time to a reasonable order of magnitude (hours instead of days).

1) Stopword vocabulary: We used the methods presented in Section III as baselines for our experiments and applied them to our datasets as if they were "regular" collections for Author Identification. These collections generally consist of long (at least 200 [18], 250 [17] or more [9] words) documents like newspaper articles [23], poems [2], letters [2], email [4] or blog posts [18]. Our collection, instead, is composed of shorter documents (see Section IV-A).

In processing the documents in our collections we follow the common practice of concatenating all texts generated by a single author together to build the profile for that author [2], [6], [23]. We indexed the author profiles using Lucene[4] with the embedded *SimpleAnalyzer* parser, which tokenises the text by using all non-letter characters as separators and lowercases it [32]. We also experimented with the other predefined Lucene analysers, however we found that *WhitespaceAnalyzer* was introducing too much noise while *StopAnalyzer* and *StandardAnalyzer* were removing the stopwords which we wanted to preserve. We also preserved the original spelling of each term, not applying any stemming at all. The choice to not remove stopwords and not perform any stemming has two justifications: i) given the nature of our dataset (i.e. composed of conversations that are user-generated), any spelling variation (including mistyping and spelling errors) can be used as a indicator of a particular user and ii) traditionally in Authorship Attribution the most frequent words (generally the majority of the terms in a standard IR stop word list) are used to identify the desired user.

Our goal with these experiments is to verify the suitability of the two statistical models, presented in Section III-A, when applied, without any specific adaptation, to the conversational documents in our collections. Each model, in fact, makes use of a dedicated set of terms as input [23] but these are mainly stopwords that have been proved to work well to profile the style of an author. With this first set of experiments we want to demonstrate that standard stopwords lists are not easily transferrable to our conversational documents and that we might need to generate a customised list of stopwords for our collections.

According to [26], a standard stopword list contains between 7 and 300 terms determined by observing the frequency of all terms in the collections and extracting those with higher frequency (we call this *Term Frequency (TF)* method). However, in [33] better results were obtained by instead considering the normalised inverse document frequency (NIDF), defined originally in [34] as $\text{NIDF}_{k \text{ Norm}} = log\left[{}^{(\text{NDoc} - \text{D}_k + 0.5)}\!/_{(\text{D}_k + 0.5)}\right]$ (we refer to this as *NIDF* method). We applied both strategies for generating the list of stop words and we also considered a standard list of stopwords taken from the INDRI search engine[5] (we refer to this as *Indri*).

In Table II we report the results of our investigation, where we show the average accuracy rate across all the partitions of the 3-fold cross validation for each stop wordlist generation strategy. We compute the accuracy rate measuring the percentage of correct assignment at the author profile level. This corresponds to the macro-average in [23]. It is worth mentioning that it is not possible to compute the micro-average in the case of our collections, because the documents are simply too short and the possibility of confusing them - due to their brevity - is too high [3]. In terms of number of users we consider 3 settings. In the first, we employed 20 users for both training and testing sets with a longer profile, emulating standard approaches [23]. In the second, we increase the number of users whilst remaining in a closed environment (same number of users for both training and

[4] Lucene version 4.0, a standard indexing and search engine library available at http://lucene.apache.org/core/

[5] http://www.lemurproject.org/

testing), however with a reduced profile length, suited for conversational documents. In the third we allow other users to be in the training set, thus having different number for training and testing set, therefore moving to the "open class" authorship attribution problem. As in the previous case, the profile length of these users is similar to conversational documents (about 40 terms.).

The choice of the number of terms to be used as stopwords was not only influenced by the standard stopwords list length [26] but also by the observations in [23] regarding the influence of the list length on the performance of the different methods. For these reasons we set the number of stopwords to a maximum of 250. Following the same standard procedures, before computing the stopword lists and computing user profiles, we removed those terms in the collections that had a document frequency of 1 and those which reoccurred fewer than 10 times in the collection.

Table II reports the results of our first set of experiments. We note that in the first setting - with 20 users and longer profiles - the stopword lists created based on the collections yielded better results compared to using a predefined list. The results obtained were expected and in-line with those in the literature for standard documents. This demonstrates that if we force our user profiles to be of an adequate length then we might achieve good results. However, this is very hard to do in practise and simply impossible for most of the users. In a real case it is not possible to determine a-priori the length of profile of a user involved in a chat and is even more difficult to force the user to produce such documents on purpose. This would be very artificial and therefore not representative of the real nature of the documents under investigation. This is evident when analysing the typical conversational documents, where the performance of the classifiers reduced dramatically.

Although the introduction of a large number of users might have influenced this behaviour, we believe the major impact in decreasing the performance has to be associated with the usage of the typical chat user's profile. The factor of influence is not only the user profile length but also its combination with the style of the documents, rendering its use as support for the classification useless. When we introduce new users into the training set and make the problem an "open class" one, we do not observe such a big drop in performance. This reinforces our belief that the explanation of the performance drop has to do with the documents' properties. We also think that making better use of the conversational nature of documents should at least partially mitigate this.

2) Single-author vocabulary: To improve the performance of the classifier and study the properties of the conversational documents, we now investigate a straightforward approach. Here we use all the terms in the profile of an author as support for the classifiers and refer to this as *simple profile*. We realised that this approach may be too simple and also computationally expensive since it involves all the terms in the profile of each user. For this reason we decide to also employ a vocabulary reduction based on the selection of the most representative terms for each user, calling it *filtered profile*. In the vocabulary reduction step for each term in the user profile we compute its probability and compare it to the probability of the corresponding term in the collection, obtaining a score. We then order terms by score and take only those with the highest scores, up to a limit to 250. Observing the results of these experiments in Table III, we notice that the performance of the *filtered profile* method is quite poor and we suspect that this might be due to the upper bound of 250 terms. We believe that this should be studied into more detail, but leave this for future work, where we want to study the influence of these thresholds on the performance of our classifiers more extensively.

Before analysing the results of the other approach, we explain two additional settings we introduce for this and following experiments. We need to introduce a minimum profile length and a minimum overlap between profile in order to make the experiments feasible. This was not needed when using the stopwords lists, because these automatically reduced the support size to at most 250 words, the maximum size of the stopwords list. When considering all the terms in the user profiles, we obtain a very large support set of terms (some thousands), even if we increased the constraints of minimum document frequency to 2 (so discarding more spelling mistakes and odd words, not really user discriminant terms). For this reason we decided to impose two additional conditions: a minimum document frequency, which we tested at 250 and 40. The first is the desiderata profile length for standard documents while the second is the average profile length for conversational documents (see Table I). We kept the minimum overlap quite low, allowing for more comparisons, and set it to between 20% and 25% of the total profile length: for a minimum profile length of 250 terms the minimum overlap limit was set to 50 and for profile of minimum 40 terms to 10. In doing so we managed to run the experiments in a reasonable time frame, without loss of generality. The first effect of these choices can be seen in the decrease in number of users in each experimental setting (except for the case where this is fixed at 20). This had an effect on the performance of the classifier, as we observed that with fewer users the performance increases, due to a reduction in the confusion between user profiles.

If we now observe the *simple profile* approach, we note a general performance improvement of the classifiers on all the experimental settings compared to the *filtered profile*. This is due to the additional information that is available from a more complete profile, rather than a filtered one. The classifier based on *KLD* seems to work better than the one based on χ^2, although this is more evident for longer profiles (min 250 terms) than for shorter ones (min 40 terms). For the same situation (longer vs. shorter profile), the dataset *irc logs* appears to be less sensitive to the minimum profile length, possibly due to the greater number of longer documents.

3) Multiple-authors vocabulary: We now look at our proposed approach, as described in Section III-B, in which we expand user profiles in the training set based on the list of

TABLE III

RESULTS FOR SIMPLE METHOD BASED ON THE CONSTRUCTION OF USER PROFILE BASED ON MESSAGE CONCATENATION

(a) User profile of at least 250 words. Comparison done if two user profiles have at least 50 overlapping terms.

Number of User		Collection	Model	Simple	Filtered
Train Set	Test Set			Profile	Profile
20	20	irc logs	kld	0.97	0.38
			chi2	0.93	0.38
		krjin	kld	1.00	0.64
			chi2	0.96	0.48
72	72	irc logs	kld	0.83	0.16
			chi2	0.67	0.16
96	96	krjin	kld	0.87	0.38
			chi2	0.60	0.35
239	72	irc logs	kld	0.73	0.12
			chi2	0.57	0.12
427	96	krjin	kld	0.73	0.19
			chi2	0.41	0.18

(b) User profile of at least 40 words. Comparison done if two user profiles have at least 10 overlapping terms.

Number of User		Collection	Model	Simple	Filtered
Train Set	Test Set			Profile	Profile
20	20	irc logs	kld	0.97	0.38
			chi2	0.93	0.38
		krjin	kld	1.00	0.50
			chi2	0.97	0.33
214	214	irc logs	kld	0.35	0.06
			chi2	0.33	0.06
707	707	krjin	kld	0.22	0.04
			chi2	0.22	0.05
1373	315	irc logs	kld	0.32	0.04
			chi2	0.30	0.05
4050	707	krjin	kld	0.18	0.04
			chi2	0.14	0.04

TABLE IV

RESULTS FOR METHOD BASED ON CONSTRUCTION OF USER PROFILES DEPENDING ON INTERLOCUTORS OF A USER. †: RESULT STATISTICAL SIGNIFICANT COMPARED TO THE SAME RESULT IN THE SIMPLE METHOD, SIMPLE PROFILE. PAIRED Z-TEST, $p < 0.05$.

(a) User profile of at least 250 words. Comparison done if two user profiles have at least 50 overlapping terms.

Number of User		Collection	Model	
Train Set	Test Set			
20	20	irc logs	kld	0.92
			chi2	0.82
		krjin	kld	0.98
			chi2	0.98†
132	132	irc logs	kld	0.52
			chi2	0.52
301	301	krjin	kld	0.89
			chi2	0.86†
668	148	irc logs	kld	0.61
			chi2	0.61
3294	343	krjin	kld	0.95†
			chi2	0.90†

(b) User profile of at least 40 words. Comparison done if two user profiles have at least 10 overlapping terms.

Number of User		Collection	Model	
Train Set	Test Set			
20	20	irc logs	kld	0.92
			chi2	0.90
		krjin	kld	0.90
			chi2	0.90
280	280	irc logs	kld	0.45
			chi2	0.46
1062	1062	krjin	kld	0.50
			chi2	0.53
1372	314	irc logs	kld	0.54†
			chi2	0.51†
6605	1198	krjin	kld	0.63†
			chi2	0.59†

each user's interlocutors. We use the same settings as in the previous experiments, having two cuts for profiles at 250 and 40 and two overlap threshold, 50 and 10 respectively.

Looking at the performance with this method in Table IV, we note a slight decrease for the controlled set of 20 users, because with so few users the benefit of incorporating the interlocutors is cancelled out. On the other hand, if we observe the "closed class" (central part of each table), we note the opposite behaviour. Performance increases for *krjin* dataset in both settings (250 and 40) with both classifiers, while this is only true for the *irc logs* for the user profile cut at 40. This is not surprising, since there are many more users in this setting, thus increasing the probability of finding interlocutors and thereby an enrichment of the vocabulary of the user. The same can be observed in the "open class" problem (bottom part of each table), where performance increases only for *krjin* in the setting of profile cut at 250 terms, while there is an improvement for both datasets when moving to the profile cut at 40 terms. Once again, this is likely explained by the increase in the number of interlocutors per user, which our methods uses to boost the classification accuracy. Before conducting further analysis we should also note an increased level of stability in the performance of both classifiers in our proposed method, where the variation is a lot less compared to the previous setting.

V. FUTURE WORK AND CONCLUSIONS

In this work we studied the problem of Authorship Attribution for conversational documents, such as IRC chats logs, and demonstrated the unsuitability of standard approaches in this setting. In particular we noted that moving from an artificial setting, such as when the number of users is heavily restricted (around 20), into the real collection with thousands of authors, the need for better adapted techniques becomes clear. For this reason we proposed a novel method that, making use of the conversational nature of our documents, is able to significantly improve the accuracy for the Authorship Attribution problem over simpler and more standard methods. Testing this method on different settings, allowing larger (250 terms) or shorter user profiles (40 terms), and with up to thousands of users in an unbalanced realistic scenario (open class problem), we were able to obtain author detection accuracy up to 95%.

These results are encouraging but also call for a more thorough investigation on the influence of the number of interlocutors in the performance gain, which we plan to study in future work. They are also very interesting in a real world context where one would like to understand the behaviour of users with lots of conversations, especially if he or she is talking with different persons, for example in the field of cybersecurity or espionage. Companies providing such IRC systems or chat services might better profile their users based on their interlocutors and more accurately address advertisement or suggestions.

Other improvements of this work in future might include the usage of use other metrics to select discriminant terms for each user, e.g χ^2 or cosine similarity, instead of KLD. It would also be interesting to try other statistical methods as classifiers, as indicated in [23], for example the Delta score or the Z score. Finally, it might also be interesting to study the influence of the length of the set of support terms in the case of simple profile study or try to combine the interlocutors effect with the temporal aspect of the chats, possibly looking also into the topical aspect of the conversations.

REFERENCES

[1] T. C. Mendenhall, "The characteristic curves of composition," *Science*, pp. 237–246, 1887.

[2] P. Juola, "Authorship Attribution," *Foundations and Trends in Information Retrieval*, pp. 233–334, 2006.

[3] G. Inches, M. J. Carman, and F. Crestani, "Investigating the Statistical Properties of User-Generated Documents," in *Proceedings of the 9th International Conference on Flexible Query Answering Systems (FAQS)*, 2011, pp. 198–209.

[4] S. Argamon and P. Juola, "Overview of the international authorship identification competition at pan-2011," in *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[5] P. Juola, "An overview of the traditional authorship attribution subtask," in *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[6] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology*, pp. 538–556, 2009.

[7] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, pp. 87–106, 1994.

[8] S. Argamon and S. Levitan, "Measuring the usefulness of function words for authorship attribution," in *In Proceedings of the 2005 ACH/ALLC Conference*, 2005.

[9] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *Journal of the American Society for Information Science and Technology*, pp. 9–26, 2009.

[10] G. Inches and F. Crestani, "Overview of the international sexual predator identification competition at pan-2012," in *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[11] R. Layton, S. McCombie, and P. Watters, "Authorship attribution of irc messages using inverse author frequency," in *3rd Cybercrime and Trustworthy Computing Workshop (CTC)*, 2012, pp. 1–8.

[12] T. Kucukyilmaz, B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," *Advances in Information Systems*, pp. 274–283, 2006.

[13] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication," *Information Processing & Management*, pp. 1448–1466, 2008.

[14] L. Wang and D. W. Oard, "Context-based message expansion for disentanglement of interleaved text conversations," in *NAACL '09*, 2009, pp. 200–208.

[15] M. Elsner and E. Charniak, "Disentangling Chat," *Computational Linguistics*, pp. 389–409, 2010.

[16] ——, "You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement," in *Proceedings of ACL-08: HLT*, no. 834–842, 2008.

[17] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, pp. 378–393, 2006.

[18] M. Koppel, J. Schler, S. Argamon, and E. Messeri, "Authorship attribution with thousands of candidate authors," in *Proceedings of ACM SIGIR conference*, 2006, pp. 659–660.

[19] R. Layton, P. Watters, and R. Dazeley, "Authorship attribution for twitter in 140 characters or less," in *2nd Cybercrime and Trustworthy Computing Workshop (CTC)*, 2010, pp. 7–13.

[20] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents (SMUC)*, 2011, pp. 37–44.

[21] G. Inches, M. J. Carman, and F. Crestani, "Statistics of online User-generated short Documents," in *Proceedings of the 32nd European Conference on IR Research (ECIR)*, 2010, pp. 649–652.

[22] R. Bache, F. Crestani, D. Canter, and D. Youngs, "A language modelling approach to linking criminal styles with offender characteristics," *Data & Knowledge Engineering*, pp. 303–315, 2010.

[23] J. Savoy, "Authorship Attribution Based on Specific Vocabulary," *ACM Transactions on Information Systems*, pp. 1–30, 2012.

[24] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," *Literary and Linguistic Computing*, pp. 251–270, 2007.

[25] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2006.

[26] P. R. Christopher D. Manning and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[27] P. Juola, "What can we do with small corpora? Document categorization via cross-entropy," in *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, 1997.

[28] ——, "Cross-entropy and linguistic typology," in *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, 1998, pp. 141–149.

[29] P. Juola and H. Baayen, "A controlled-corpus experiment in authorship identification by cross-entropy," in *Literary and Linguistic Computing*, 2003, pp. 59–67.

[30] Y. Zhao, J. Zobel, and P. Vines, "Using relative entropy for authorship attribution," in *Proceedings of the 3rd Asia conference on Information Retrieval Technology*, 2006, pp. 92–105.

[31] D. Jurafsky and J. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall/Pearson education international, 2008.

[32] E. Hatcher and O. Gospodnetic, *Lucene in Action (In Action Series)*. Manning Publications Co., 2004.

[33] R. T.-W. Lo, B. He, and I. Ounis, "Automatically building a stopword list for an information retrieval system," *Journal of Digital Information Management*, pp. 3–8, 2005.

[34] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information Science*, pp. 129–146, 1976.

APPENDIX

A. Channels in "irc-log"

aix apache azureus blender c cisco csharp css debian fedora flood freebsd gentoo gentoo-dev gtk hardware html iptables irix java javascript linux-bg macosx mysql netbsd openbsd opensolaris oracle php python qt reactos samba solaris suse tomcat ubuntu vim windows wireless

B. Channels in "krijn'

accessibility activity css developers fx html-wg html5 microformats wai-aria webapps whatwg xhtml