A Feature Selection Method for Author Identification in Interactive Communications based on Supervised Learning and Language Typicality

Esther Villar-Rodriguez [a], Javier Del Ser [a,b,c,*], Miren Nekane Bilbao [b] and Sancho Salcedo-Sanz [d]

[a] *OPTIMA Area, TECNALIA, 48160 Derio, Bizkaia, Spain.*

[b] *Department of Communications Engineering, University of the Basque Country EHU/UPV, 48013 Bilbao, Bizkaia, Spain.*

[c] *Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain.*

[d] *Department of Signal Theory and Communications, Universidad de Alcalá de Henares, 28871 Alcalá de Henares, Madrid, Spain.*

Abstract

Authorship attribution, conceived as the identification of the origin of a text between different authors, has been a very active area of research in the scientific community mainly supported by advances in Natural Language Processing (NLP), machine learning and Computational Intelligence. This paradigm has been mostly addressed from a literary perspective, aiming at identifying the stylometric features and writeprints which unequivocally typify the writer patterns and allow their unique identification. On the other hand, the upsurge of social networking platforms and interactive messaging have undoubtedly made the anonymous expression of feelings, the sharing of experiences and social relationships much easier than in other traditional communication media. Unfortunately, the popularity of such communities and the virtual identification of their users deploy a rich substrate for cybercrimes against unsuspecting victims and other forms of illegal uses of social networks that call for the activity tracing of accounts. In the context of one-to-one communications this manuscript postulates the identification of the sender of a message as a useful approach to detect impersonation attacks in interactive communication scenarios. In particular this work proposes to select linguistic features extracted from messages via NLP techniques by means of a novel feature selection algorithm based on the dissociation between essential traits of the sender and receiver influences. The performance and computational efficiency of different supervised learning models when incorporating the proposed feature selection method is shown to be promising with real SMS data in terms of identification accuracy, and paves the way towards future research lines focused on applying the concept of language typicality in the discourse analysis field.

Key words: Authorship Identification; Natural Language Processing; Supervised

* Corresponding author: Javier Del Ser. OPTIMA Area. TECNALIA, 48160 Derio, Bizkaia, Spain. T: +34 946 430 850. Fax: +34 944 041 445
Email address: javier.delser@tecnalia.com (Javier Del Ser).

1 Introduction

Authorship attribution refers to the discipline aimed at distinguishing the author or writer of a certain text by processing and analyzing features extracted from the content under consideration. Although early studies in this field date back to more than 100 years ago, this area has undergone a sharp activity increase in the last decade as the result of several advances in machine learning and Natural Language Processing (NLP), further ignited by new data management and processing trends under the so-called Big Data paradigm. Similarly, related subareas such as author profiling and authorship deception have furnished the literature with a plethora of contributions dealing with the application of supervised models to these problems. Disregarding the ultimate aim of such contributions, most of the works reported to date resort to similar text representation strategies, writer-specific features and classification models, mostly at the pace dictated by the progress in textual information retrieval and Computational Intelligence (e.g. see [1,2] and references therein).

Traditionally the scope of authorship attribution has been mostly focused on long pieces of text delivered over unidirectional and/or non-interactive communication means (e.g. from books to letters and research articles), which intuitively implies a stylistic content homogeneity. However, the advent of more dynamic messaging applications such as the Short Message Service (SMS), chats, micro-blogging and social networks has given rise to recent experiments over these interactive, bidirectional channels [3–5]. All contributions tackling this particular communication scenario state that short-length textual contents encompass strong technical challenges due to the shortage and low diversity of vocabulary that determine the predictive richness and uniqueness of the extracted features and ultimately, the accuracy of subsequent classification models [6,7].

Vocabulary richness can be indeed measured at different levels. To begin with, stylometry refers to the application of linguistic and tonal style analysis striving to unambiguously identify the writer. According to [8], stylistic features include a vast number of lexical, syntactic, structural, content-specific, and idiosyncratic style markers. Despite its inherent computational complexity and reduced scalability, off-the-shelf text parsing can exploit lexical, token-based and syntactic features, yet they result in a high variance and inaccuracy when dealing with undersized text samples. Vocabulary richness can be also measured by counting either unique terms or those word forms appearing exclusively once or twice in the entire text under analysis (i.e. the so-called *Hapax* and *Dis Legomena*), which can be often approximated by a Zipf distribution [9]. Other measures of vocabulary richness have been extensively used as word-

base methods to characterize the variance and diversity of a given glossary [1]. Measures such as Yule's, Simpson's, Honoré's, Sichel's or Brunet's are devised to diverge from prior biased procedures (e.g. the archetypical type/token ratio measure) in relation to the text size [11]. Nonetheless, when dealing with short texts in non-formal contexts a convenient option to characterize the lexical register is to adopt the character n-grams scheme by representing the elocution in n-sized splits of letters, being this at the same time a robust and noise-tolerant artifice to contend with spelling errors.

Furthermore, the morphology of the written record may be justified by the evaluation of the structural aspect of language. Syntactic features comprise sentence- and word-length distribution, the use of pronouns, conjunctions or other parts-of speech of interest, which reveal the developed and desired structure to heighten aspects such as the connections amongst the constituents, subject and clause shifts and broadly speaking the mixture of diction and grammatical complexity. Parts-of-Speech (PoS) tagging groups together those words with similar grammatical function (e.g. noun, adjective or verb), being usually combined in sets of tags or constituents according to their syntactic function so as to register the discourse assemblage [6]. Other novel approaches have opted for gathering and arranging them into rewrite rules showing, in terms of frequency, the hierarchy and composition preferences of the author [12]. On the other hand, connective particles, as conversational connections, reflect the nature of the segmentation from a structural or grammatical view, as well as the writer temper and inclinations [13].

In line with the vocabulary richness and uniqueness required for authorship identification, it is important to note that certain grammatical categories empower the meaning and the semantics conveyed inside the prose specifying the attitude or mood of the author, whereas others are typically structural and bear little significance or connotation. Topical elaboration is well represented by adverbs and adverbial expressions, whereas slightly appreciable semantic information can stem from 1) the tense and aspect of the employed verbs; 2) the relations between a PoS node and its children in the parsing tree [14]; and 3) the use of lexical resources such as synonyms or hyperonyms, which are used whenever ambiguity or abstraction is required [15]. However, not only strict and order-driven grammar must be captured by language parsers, but also phrase and predicate-argument structures can be functionally inferred by deep parsers to avoid wrong PoS miss-classifications when facing the recurrent ambiguity of attachment in long-distance relationships [16].

As mentioned before, most literature resorting to the above measures of linguistic richness for authorship attribution has focused on literary records rather than interactive one-to-one communications, in which the speech is

[1] The authors refer to [10] for a more thorough description.

generated in a dynamic basis with a two-way flow of information. In this alternative communication scenario the majority of the previously surveyed measures are deemed short and insufficient in terms of prediction accuracy. The rationale behind this statement lies in the intuition that in such a dyadic communication the noise factor distorts the message by introducing elements such as inattention, disinterest or cultural differences, which finally produce as many diverse dialogues as different receivers are involved. These multiple channels subject to diverse contexts and influences call for an independent analysis of their heterogeneity in an attempt at preventing authorship attribution models from counting on exceptional, occasional or receiver-influenced linguistic features.

When approaching authorship identification from a machine learning perspective, the ideal scenario is that where instances (i.e. texts) belonging to the same category (correspondingly, author) are confined within compact clusters in the space spanned by the utilized features. In such an idealized setup clusters should be restricted to the feature *essence* of the writer so as to avoid building strongly adjusted predictive models capable of capturing multiple yet infrequent patterns of the exchanged messages. This may eventually lead to overfitting and consequently, to a poor predictive performance in terms of the generalization properties of the classifier. This manuscript gravitates precisely on how to isolate in practice the feature essence of the writer from the context and the influence of the receiver on the message so as to exploit it in authorship attribution. This hypothesis has a particular application focus on impersonation and identity theft attacks in social networks; previous contributions have commonly highlighted the increasing incidence and severity of this class of subtle cybercrimes, particularly within the teenage community [17,18]. The requirements posed on an impersonation detector for one-to-one communications in terms of scalability and sender diversity may find a suited technological response in the concept of feature essence tackled in this manuscript.

An example may clarify the above application of the proposed methodology to the detection of impersonation in social networks. In the scenario depicted in Figure 1 (left) Bob establishes several chats with his friends Alice, Frank, Craig, Grace and Carol. The social network platform records the messages sent by Bob and extracts, by virtue of the proposed methodology, a set of essential linguistic features that characterizes Bob when chatting with his friends. This feature set is the basis for training a supervised model (in general, one model for each user of the social network) with messages verified beforehand to belong to the real Bob. Eventually (Figure 1, right) a hacker (denoted as *other* Bob) impersonates the legitimate Bob and starts communicating with Bob's social circle as if he/she were actually Bob. The model trained in the previous stage declares that the essential features of the messages sent by the *other* Bob do not follow the pattern of the legitimate Bob, hence the social net-

work administrator labels such messages as *suspicious*. This methodology can be adopted not only to unveil malicious or impersonated uses of a legitimate account (Bob's), but also to reveal different roles of the sender. This latter purpose unleashes several applications of interest if the essential feature commonality is further explored across confirmed cases of dual identities played from a single account, such as the identification of pedophiles, terrorists and other profiles/roles alike.

An experimental setup has been designed to explore our hypothesized concept of feature essence, its impact on an authorship attribution scenario and the universality of its performance when applied over supervised learning models of very distinct nature. Due to the lack of public datasets with confirmed cases of impersonation, experiments will analyze the benefits of our proposed feature selection approach when applied to the identification of the sender of a certain message among a fixed set of authors. While the impersonation detection scenario exemplified in Figure 1 would require one-class classifiers for its implementation, the analysis of the proposed feature selector when applied to a multi-class authorship attribution problem allows for a better performance characterization of the supervised learner to which the selected features are fed. Furthermore, we avoid any need for characterizing the feature space of the *other* user (i.e. the counter-examples not present in the training set).

At this point it is relevant to recall that our hypothesis postulates that content features coming from distinct yet distinguishable senders allow for a new feature selection criterion based on isolating the linguistic pattern of the sender that is invariant with respect to the receivers. This essential feature set is expected to impact positively on the subsequent authorship attribution task in terms of 1) the scalability and computational complexity of the utilized classifier, due to a vastly reduced feature space; and 2) the generalization properties of the model as the number of users grows, with more relevant, predictive characteristics being fed to its learning procedure. Simulation results will indeed underpin that the proposed technique renders good prediction scores when compared to other off-the-shelf feature selection methods for a given supervised learner.

The rest of the manuscript is structured as follows. First, Section 2 describes the proposed feature selection scheme, whereas Section 3 elaborates on the details of the data set utilized for assessing its performance. The experimental setup and the results obtained therefrom are outlined and discussed in Sections 4 and 5. Finally, Section 6 concludes the paper and sketches future research.

2 Proposed Feature Selection Approach

In formal contexts, singularities are often derived from the frequency of word, n-grams or syntactic elements considered as specific author's stylistic choices. The variance or the information contained in such distinctive elements determines the precision in the identification of the sender. Nevertheless, in more dynamic environments as the one considered in this paper other characteristics must be addressed such as the usage of emoticons and/or punctuation marks, which usually evinces the emphasis or the intensity of the text. Based on this rationale, a selected set of features has been assembled so as to compile the linguistic peculiarities of every sender within a diversity of contextual communication scenarios. The overall set of linguistic predictors comprises the following items:

(1) Word-based features: the word length distribution and the character trigrams.
(2) Grammatical features: adverbs, adjective and first-person frequency.
(3) Syntactic features: PoS bigrams, sentence complexity (measured as the number of composite – coordinated or subordinate conjunctions – clauses) and function words distribution.
(4) Social media and instant messaging based features: punctuation, distribution of emoticons and slang abbreviations (own compiled dictionaries [2] with 181 and 1137 regular expressions, correspondingly).

Before any further grammatical or syntactic processing, trivial procedures have been applied to normalize the messages within the dataset: removal of uppercase letters [3], repeated characters and slang abbreviations, the latter after annotation [4]. The word length embraces the so-called concept of readability as a text-inherent factor quantifying the lexical involution. The syntactic features have been extracted by means of a model trained on Twitter driven by the Stanford PoS Tagger [19,20], which allows for a sophisticated treatment of the inconsistency and ungrammaticality of the messages within this particular dataset. The analysis of adverbs and adjective usage represents the topic elaboration and the grade of quality description implemented by the sender. In turn, sentence complexity refers to the tendency to construct subordinated or dependent and coordinated phrases, which implicitly quantifies the complexity of the syntax structure of the message at hand.

[2] Resources utilized in this paper will be made available to the community in a public repository once it is published.
[3] Combinations of upper and lowercase letters in the words could be also explored as potentially essential predictors.
[4] This annotated slang corpus can be made available on demand.

Once these features have been computed and collected for each message, it should be noted that their cardinality might increase with the number of distinct senders. For instance, the number of different trigrams compiled over the dataset depends on the diversity and similarity of the messages exchanged among different sender-receiver pairs, and is closely linked to the concept of essence postulated in this work. Indeed, an empirical analysis of the total number of features, PoS bigrams and trigrams reveals that they all grow as the number of senders increases. However, as shown in Figure 2 differences are minimal when adding the features corresponding to the sixth within the selected set of senders. Interestingly this manifests the fact that in a dynamic corpus with short messages as the one utilized in this paper, linguistics are more likely to become homogeneous and less diverse. Nevertheless, from the plot it should be inferred that when determining whether any given message corresponds to a given sender, any criterion focused on reducing the overall number of features should be of interest in order to avoid subsequently overfitted classification models and decrease the computational complexity of their training process.

When dealing with classification tasks, information gain, odds ratios or tests such as the Kullback-Leibler divergence [21] or Chi-Squared [22] are widely applied in the search for the most discriminatory predictors. Nevertheless, these typical feature selection algorithms are exploited as an early and independent stage and regardless the context or the problem at hand. In this work we propose a rather different feature selection algorithm well-suited for multi-class authorship attribution models composed by independent OvO (One Versus One) classifiers.

For the sake of understandability, in what follows *essential* and *influential* features will stand for the selected feature subsets by the proposed technique, which springs from theoretical concepts of linguistics. When aggregating features from dyadic dialogues in an attempt to discern the sender of a message a numerous of sporadic, context-dependent linguistic elements can be captured, which are likely to generate over-sized collections of features and potentially overfitted classification models. This expansion phenomenon will become sharper when dealing with hundreds of senders and thousands of messages as in dyadic communications over social networks; the detection of impersonation attacks in this scenario requires a fine-grained characterization of the linguistic feature essence of each user in order to avoid very intricate decision regions for the classifier that might lead to a high rate of false positives. We assume that essential patterns can be more productive in the long run when considering several aspects:

- The instability was a criterion for feature selection introduced in [23] under which those terms that can be changed by other alternative terms (synonyms) conform the stylistic choices of the author, as opposed to those ones

of forced usage such as some prepositions or monosemic lexicon with no feasible variants. In a purely literary context the author is willing to tune his/her content as much as possible. However, in an informal communication context as the one held through social networks nearly the opposite approach holds: authors do not elaborate on the opinions nor use complicated language to polish the message content. In this scenario frequent discourse choices are broadly representative of the sender, who may select terms unconsciously across any possible destination thus becoming advantageous to discern among different authors.

- As the number of users grows, so does the cardinality of the feature set and potentially, the amount of behavioral patterns that must be discerned by the classifier. This gives rise to a higher complexity of the model due to the need for partitioning regions in the feature space that lead to well-generalized decisions in regards to the authorship of the messages under test. Furthermore, it should be assumed that the short average length of the messages and their usage context could unavoidably homogenize their features and consequently, imply a loss of predictability in regards to their authorship that cannot be overridden (not even by a social psychology specialist). In addition, we cannot expect linear patterns related to topics for a specific individual over the time; conversations through the considered communication channels use to be more topically diverging from each other, often implemented over more diverse vocabulary than in books, articles and more static media.

Based on the above two observations, a more flexible feature selection algorithm has been devised. It should be again emphasized that the proposed scheme can be applied to both one-class and multi-class authorship attribution models. While the former corresponds to the detection of impersonation attacks in social networks, the latter is deemed appropriate for the characteristics of the selected dataset. When dealing with one-class classifiers the performance assessment usually becomes more involved than the multi-class set, in part due to the lack of ground of truth to which to benchmark the obtained predictive outcomes. This being said, the derived feature selection method is hereafter contextualized and put to practice over a multi-label classification scenario where the sender for the tested SMS's must be discriminated. Nevertheless, discussions will be held on the extrapolation of the conclusions extracted therefrom to the one-class case.

The feature selection scheme proposed in this manuscript aims at isolating the feature essence of each user, and exploiting the union of the essential feature sets of the pair of senders involved in each OvO classifier. In reference to Figure 3, two approaches can be arranged in this envisaged application scenario:

- Approach A: this corresponds to the naive concatenation of a feature selection (FS) stage and a multiclass classifier. The preprocessing stage is in

charge of discriminating the most predictive features from the overall set of extracted characteristics. As aforementioned in the introduction, this can be performed in very diverse ways. The approach is completed by a multiclass classifier, which may be implemented by resorting to any supervised learning model.

- Approach B: now the feature selection algorithm is split in several stages FS_i (one per sender), each in charge of collecting all samples in their original size sent by user i, and selecting exclusively those features (essence) that are used recurrently along the entire set of messages sent by sender i to any receiver. Then a multiclass classifier is built by deploying $S(S-1)/2$ OvO classifiers (with S denoting the overall number of distinct senders), each fed with the union of the essential feature set of the users to be distinguished. The final decision results from voting the outputs of the OvO classifiers. The intuition behind this approach resides in the hypothesis that the essence of a sender consists of those less linguistic singularities that hold in every communication between him/her and any third party. Consequently, these essential features are invariant and are not affected by the influence of any receiver or the context, so they should remain present in future messages sent by the same sender. Therefore, the proposed procedure first splits the message set of each sender in disjoint sets depending on the receiver to whom they are sent, and next computes a occurrence frequency histogram of each feature over each of such subsets. The essential set for such a sender-receiver pair results from discarding those features whose frequency of occurrence falls below a given threshold. These selected features belonging to a certain sender-receiver conversation are then intersected with the rest of the filtered sender-receiver feature subsets on the basis that those commonly shared features delimit the interlocutor essence and are context-insensitive. In other words: the imposed threshold selects the most recurrently used features over communications held between a given sender and his/her different receivers, whereas the intersection of such feature sets defines all such characteristics that are recurrently used by the sender at hand independently of the receiver or the context.

In mathematical terms and in reference to Figure 3, let the k-th message from sender i to receiver $j \in \mathcal{J}_i$ be represented by $\mathbf{m}_{i,j}^k \doteq \{m_{i,j}^{k,0}, \ldots, m_{i,j}^{k,N-1}\}$, with $\mathcal{N}$ denoting the set of originally extracted features with cardinality $N = |\mathcal{N}|$, and $\mathcal{J}_i$ the set of receivers of sender i. The purpose of the feature selection algorithm is to compute a feature subset $\mathcal{N}_i \subseteq \mathcal{N}$ such that given two different senders i and i', the union set $\mathcal{N}_{i \cup i'} \doteq \mathcal{N}_i \cup \mathcal{N}_{i'}$ can be used at the OvO classifier yielding a better predictive performance and/or lower computational complexity by virtue of its reduced size. To this end, a N-sized vector $\mathbf{f}_{i,j} \doteq \{f_{i,j}^n\}_{n=0}^{N-1}$ containing the frequency of occurrence of each feature between each

sender-receiver pair (i, j) is computed as

$$f_{i,j}^{n} \doteq \frac{\sum_{k=1}^{K_{i,j}} \mathbb{I}(m_{i,j}^{k,n} > 0)}{K_{i,j}}, \tag{1}$$

where $K_{i,j}$ represents the number of messages from sender i to receiver j, and $\mathbb{I}$ is an indicator function taking value 1 if its argument is true and 0 otherwise. Once this vector has been computed, a minimum frequency threshold $\Psi_{i,j}$ determines the number of features to be retained for the sender-receiver pair as

$$\mathcal{N}_{i,j} = \left\{ n \in \mathcal{N} \ : \ f_{i,j}^{n} \geq \Psi_{i,j} \right\}, \tag{2}$$

from which the set of essential features for sender i is given by

$$\mathcal{N}_{i} = \bigcap_{j \in \mathcal{J}_i} \mathcal{N}_{i,j}. \tag{3}$$

From the above formulae and Figure 4 it should be obvious that $\Psi_{i,j}$ plays a crucial role in determining the minimum occurrence support that a feature should meet to be essential in the communication channel between sender i and receiver j and, eventually, for the sender i at hand if such an essential nature holds when assessed over all his/her receivers. This threshold should be adapted to the particular occurrence profile of the features over the different communication channels of the sender. In other words, it should capture potential inflection points within an ordered occurrence histogram beyond which the remaining feature subset becomes almost uniform hence statistically irrelevant for subsequent classification tasks.

A stand-alone, self-adjusting method to detect this point $n_{i,j}^{*}$ starts by sorting $\mathbf{f}_{i,j}$ by index n in decreasing order, yielding an index mapping $\lambda : \mathcal{N} \rightarrow \mathcal{N}$. By defining points $\mathbf{p}_0 = [0, f_{i,j}^{\lambda(0)}]$, $\mathbf{p}_{N-1} = [N-1, f_{i,j}^{\lambda(N-1)}]$ and $\mathbf{p}_n = [n, f_{i,j}^{\lambda(n)}]$, the inflection point $n_{i,j}^{*}$ is given by

$$n_{i,j}^{*} = \lambda^{-1} \left(\underset{n \in \{1,\ldots,N-2\}}{\arg\min} \arccos \frac{(\mathbf{p}_0 - \mathbf{p}_n) \times (\mathbf{p}_{N-1} - \mathbf{p}_n)}{|\mathbf{p}_0 - \mathbf{p}_n| \cdot |\mathbf{p}_{N-1} - \mathbf{p}_n|} \right) \tag{4}$$

where $\times$ denotes vectorial dot product and $|\cdot|$ stands for L_2-norm. The above expression seeks the point from which the angle between the straight lines connecting points $\mathbf{p}_0$, $\mathbf{p}_{N-1}$ and $\mathbf{p}_n$ is minimum. Once this point has been computed, the set of essential features in the messages between sender i and receiver j is given by

$$\mathcal{N}_{i,j} = \left\{ n \in \mathcal{N} \ : \ f_{i,j}^{n} \geq f_{i,j}^{n_{i,j}^{*}} \right\}, \tag{5}$$

from which the essential feature set for sender i is inferred by resorting to Expression (3). This computation of $n_{i,j}^{*}$ allows for a higher flexibility and

adaptability of the threshold $\Psi_{i,j}$; in fact, it should be clear that $\Psi_{i,j} = f_{i,j}^{n^*_{i,j}}$.

3 Dataset

We begin the discussion by delving into the dataset utilized within the experiments. The selection was based on the main premise that interactive communication channels are more likely to develop contextual and receiver influences onto the exchanged messages. Intuitively short messages lay the foundation of the dyadic and interactive discourse, being thus a suitable corpus for analyzing senders' linguistics when communicating to multiple receivers. This motivates the myriad of datasets chosen in previous contributions, mostly encompassing extracts from newspapers [13,24] or books [6]. Other data sources have been lately explored as a result of the growing proliferation of Social Media, which fosters the creation and exchange of user-generated content via new highly interactive channels such as blogs [14], public on-line forums or message boards [7,16] IRC chatting systems [3,4,7] or micro-blogging environments [25].

Unfortunately, to the best of the authors' knowledge no dataset containing messages retrieved from social networking platforms is publicly available in the Internet. However, it is important to note that nowadays user habits in terms of social media have evolved towards an ubiquitous usage in mobile phones. This is especially frequent within teenagers, which gets even more usual by the latest proliferation of interactive communication means (e.g. chat) embedded in the application itself. As a result linguistics in Social Media have progressively converged to those of traditional schemes. Thereby, for our experiments we have opted for the NUS SMS Dataset [26], which is a collection of 65296 English SMS messages compiled by researchers from the School of Computing of the National University of Singapore between 2011 and 2014. Among all 65 senders within this dataset, those with at least 4 receivers with more than 100 messages have been selected for the experimental phase, accounting for a total of 13036 messages. First a minimum of 100 messages was imposed between every sender-receiver pair so as to ensure enough data to characterize the linguistic usage in the communication process, yielding a total of 27 eligible senders. Out of them only 6 users met the requirements of at least 4 different receivers. This filtering permits analyzing the concept of essence posed in this manuscript without any eventual side effect due to a low number of messages and/or receiver diversity. It should be also emphasized that to the knowledge of the authors, no other contribution has been previously made with this specific dataset apart from spam filtering (see e.g. [27–29]).

We further argue that the utilization of the NUS SMS dataset does not conflict with the particularities of message exchanging in social networks on which the scope of this work is framed: despite the fact that one-to-many channels are

allowed in such networks, it is in personal communications with different receivers where the discrimination of the linguistic essence of the sender can be performed. In other words, masqueraders, pedophiles and other attacks alike do share the same detection goal: to verify whether the sender of the message can be discriminated efficiently by leveraging linguistic features that cannot be consciously modified by the sender (*essence*). The isolation of the linguistic typicality must be performed over different one-to-one communication channels sharing the same transmitter, disregarding whether we deal with chats embedded in social networks, SMS messages or any other means.

4 Experimental Setup

Several experiments have been performed to assess the impact of our proposed feature selection scheme on the accuracy of supervised learning models when applied to the selected dataset. To this end, two different machine learning models will be utilized for implementing the multi-class (approach A) and OvO (approach B) classifiers:

- A Support Vector Machine (SVM), which constructs a hyperplane – or set of hyperplanes – in a feature space of increased (if required) dimensionality so as to map different category instances into maximally separated decision regions [31]. By maximizing the margin between the closest points belonging to different categories the generalization error is minimized. SVM permits working on features that when combined or rendered onto a larger space, become significant and decisive in terms of predictive significance. In these models the trade-off between generalization and overfitting is mainly controlled by the penalty parameter C: the larger C is, the less the final training error will be, but a higher risk is assumed to jeopardize the generalization properties of the classifier.
- A Random Forest (RF) classifier, which is built by an ensemble of simple decision trees each trained with a bootstrap sample drawn from the overall training set [30]. In addition, the split in such compounding trees is not decided among all features, but instead among a random feature subset. This randomness involves a slightly increased bias with respect to a non-randomized single decision tree. However the variance decreases more significantly to usually compensate for the increase in bias, ultimately yielding a model with enhanced predictive generalization capabilities. Random Forests are a response to those classifiers which tend to generalize without ruling out outliers or noisy patterns eventually creating models with high variance and then sensible to minor fluctuations in the training set.

The experiments discussed in what follows focus on the *universality* of the proposed technique, i.e. on verifying whether the derived feature selection

scheme is beneficial for different pattern recognition models disregarding their internal learning procedure. SVM models have been extensively utilized in the literature and proven to be flexible enough so as to process language processing datasets of high dimensionality, whereas RF models incorporate an embedded feature selection method in the construction of their constituent learners that, when trained over bootstrapped samples and bagged altogether, has been found to provide a low variance in their output. This sought diversity is the reason why both models have been included in this benchmark, not for comparing one to each other but for quantifying to what extent the proposed feature selection approach benefits each of them.

Based on this intended scope and for the sake of fairness, the classification model in Approach A will implement a OvO ensemble with the baseline classifier at hand (SVM or RF) as the constituent learning model. This will allow comparing both approaches under the same model configuration, hence minimizing any eventual influence due to differing ensembles. In regards to approach B and denoting as $D^k_{i,j}$ the authorship decision made for message k in the OvO classifier deciding between senders $i \in \{1, \ldots, S\}$ and $j \in \{i+1, S\}$, two voting schemes will be considered:

- Hard voting: the final decision about the authorship of each message is the most frequent value over the outputs $\{\{D_{i,j}\}_{i=1}^{S}\}_{j=i+1}^{S}$ of the OvO classifiers.
- Soft voting: the final decision is furnished by fusing the likelihoods produced for each sender by every OvO classifier. In the case of RF the predicted probabilities of an input sample for a certain class (i.e. sender) is estimated by averaging the predicted class probabilities of the trees in the forest, where the class probability of a single tree is given by the fraction of samples of the same class in a leaf. As for the SVM classifier, Platt scaling [32] is utilized for transforming the hard output of the model into a distribution of probabilities over classes. If $p(i|m_k)$ denotes the overall likelihood about the authorship of sender i estimated for message k, by assuming conditional independence between the OvO classifiers and uniformity among the senders when authoring the message it can be proven that the probability of sender i authoring message m_k is proportional to the product of the a posteriori likelihoods generated by the $S(S-1)/2$ OvO classifiers. In mathematical terms, by defining such a posteriori likelihood for OvO classifier $s \in \{1, \ldots, S(S-1)/2\}$ as $p_s(i|m_k)$, one obtains that

$$p(i|m_k) \propto \prod_{s=1}^{S(S-1)/2} p_s(i|m_k), \tag{6}$$

  i.e. the product of the output probabilities of those OvO classifiers where the authorship of sender i is compared to every other sender. The soft voter will opt for the sender i^* with the highest total likelihood among all possible

senders, i.e.

$$i^* = \underset{i \in \{1,\ldots,S\}}{\arg\max}\, p(i|m_k). \tag{7}$$

In summary, the methodology followed to perform the experiments and obtain the results discussed in the next section can be summarized as follows:

(1) Process the NUS SMS Dataset and select those senders with at least 4 receivers, each receiving more than 100 messages. This produces the baseline dataset with 6 senders utilized in subsequent steps.
(2) Extract the set of *essential features* for each sender in the reduced dataset by processing all messages exchanged with his/her receivers through the proposed feature selection approach in Section 2 and Expressions (1) to (5). Depending on the characteristics of the threshold $\Psi_{i,j}$ used in Expression (2) this step gives rise to essential feature sets for every sender based on *fixed* or *adaptive* thresholds.
(3) Evaluate the cross-validated performance of Approach A with multi-class OvO SVM/RF models combined with a portfolio of alternative feature selection and dimensionality reduction techniques. Record the performance statistics (mean and standard deviation) of the classification accuracy.
(4) Compute the same performance indicators for Approach B with the essential feature sets extracted for every sender with *fixed* or *adaptive* thresholds and *hard* or *soft* voting.
(5) Compare and discuss the obtained results.

5 Results and Discussion

We begin the discussion by analyzing Table 1, where results in terms of classification accuracy are shown for different models (RF and SVM) and classification approaches as previously depicted in Figure 3:

- Approach A with different feature selection schemes: variance thresholding (i.e. those features whose standard deviation across samples is zero are discarded), selection of the set of best features based on univariate statistics (χ^2 and the ANOVA F-value), average-based feature selection based on importance weights as provided by tree-based models (only features whose importance when training a RF model is greater or equal than the mean of all feature importance values are kept), Recursive Feature Elimination and the Fast Correlation Based Filter method proposed in [33]. Furthermore, linear dimensionality reduction based on Principal Component Analysis (PCA) has been also considered with different output linear components.
- Approach B with the proposed linguistic essence isolating scheme, hard/soft voting and fixed/adaptive thresholds.

The performance of each scheme under comparison has been averaged over 10 stratified folds to assess the statistical stability of the score. Furthermore, parameters controlling the utilized models (e.g. C and γ for the SVM) have been all optimized via grid search and a local 5-fold cross-validation.

As shown in this table, both models benefit from the application of the proposed feature selection technique, but in rather different terms: to begin with, SVM in Approach A with variance thresholding outperforms slightly Approach B with soft voting and self-adjusted threshold in terms of the attained accuracy score. However, SVM does benefit in terms of computational complexity, as good scores (in a comparable order to those attained with all features) are obtained by solely resorting to a very reduced subset of the overall number of characteristics, as later discussions will clearly show. This renders a noticeable difference in terms of training runtime of the model. Furthermore, a deeper analysis revealed a subset of 6132 features utilized just by one of the authors within a dataset comprising 13036 messages, which results from the combined effect of the small amount of authors involved in the experiments and the reduced size of the processed SMS texts. This uniqueness entails very distinguishable regions in the feature space containing just one single message; in this situation the particular split criteria of tree classifiers beneath RF models (e.g. GINI impurity) would not take advantage due to the lack of information gain in terms of discriminative capacity, nor our proposed feature selection would select such one-message features as they are not recurrently employed across all receivers.

In fact, by removing such unique samples an alternative setup can be emulated with a high number of users with very unlikely one-term linguistic uniqueness. In this scenario the accuracy figures of the SVM-based classifier degrade significantly (e.g. 0.605 / 0.002 for Approach A with variable thresholding), while those of its RF-based counterpart remain in the same order (correspondingly, 0.603 / 0.009). Remarkably, when removing such unique characteristics the proposed essence-based feature selection scheme provided performance gains in a similar manner to what is obtained for the RF model over the whole dataset (e.g. 0.615 / 0.003 for Approach B with hard, self-adjusted $\Psi_{i,j}$). This deeper analysis unveils that the performance gaps among supervised learning models are due to the particular statistical properties of the dataset rather than on an apparent lack of universality of our proposed method. Discussions will be hereafter held over the results obtained for the reduced version of the dataset published in [26], including the detected set of unique one-message features.

Following the above rationale, the performance gained for RF resides not only in the lower training complexity yielded by a significantly reduced feature space, but also in the performance score of the overall classifier. The reason being that the particular training algorithm of RF incorporates an embedded

random feature selection scheme at each of its compounding tree learners. Results in Table 1 suggest that the random selection of variables in the split of the constituent tree models of RF benefits from a previous discarding of all non-essential features, due to the fact that the likelihood to perform a split over a relevant variable in terms of authorship discrimination increases. Furthermore, for a given supervised learning model (either SVM or RF) our proposed feature discrimination method with soft voting is able to better isolate relevant predictors for authorship attribution than the rest of feature selectors considered in the benchmark.

As previously mentioned, the performance of the proposed approach must be further analyzed from the perspective of the computational complexity of the model training, which relates directly to the number of input features. Table 2 shows the absolute number of essential features obtained for each threshold selection method in Approach B, and their relative percentage with respect to the overall number of predictors utilized by Approach A after variance thresholding. It can be noticed that Approach B vastly reduces the number of utilized features (in particular, at most 1.51% for self-adjusted thresholding). This observation buttresses the intuition that the proposed essential feature detector not only reduces the computational complexity of the training process for SVM models at comparable performance scores, but also helps the inherent feature selection method of RF models in the discrimination of good predictors, to the point of achieving better prediction results. This conclusion gets further reinforced by assessing the computation time taken by each of the above schemes: when implemented in Python 2.7.6 on a Pentium Core i7 Pro with 16 Gigabytes of RAM, the execution of each fold in Approach A (after variance thresholding) takes on average 310.4 times longer than the longest variant of Approach B (self-adjusted $\Psi_{i,j}$). Finally, this table also encloses the overall and unique number of features handled by the constituent OvO classifiers of the different Approach B variants, the latter serving as a baseline information to set a comparable range of output components by the rest of feature selectors within the previous benchmark.

Further interesting observations can be drawn if the average precision scores shown in Table 1 are broken down into the individual metrics attained by each of the compounding OvO classifier. Even though similar conclusions can be reached from the rest of models and variants of the proposed feature selection scheme, for clarity we will focus on Approach B with RF as the core learning model and self-adjusted essence feature selection. Table 3 depicts the confusion matrix of the overall approach. Therein it can be noticed that while the precision for users 0 to 3 are very satisfactory bearing in mind the short length and limited content of the processed messages (with user 3 amounting up to a precision of 87%), confusion appears between users 4 and 5. This bad classification result is supported in part by Figure 5, which depicts the accumulated number of features (discriminated by type of feature) when the

number of users increases. This plot suggests that when considering the last user jointly with the rest of possible authors an upper bound in the number of total features can be achieved. In other words, this unveils a *linguistic* limit of the SMS messages contained in the dataset under consideration: when dealing with short-length SMS messages it is very likely that a high fraction of them shares the same n-gram set. This implies that for certain users it becomes necessary to resort to 1) information of other nature so as to uniquely identify their authorship, such as the connection usage approach proposed in [18]; and/or 2) the joint processing of successive decision outcomes along time, as argued in what follows. Nevertheless, before closing this discussion it is necessary to recall that the final goal of this research work goes beyond authorship attribution and aims at the detection of different roles in the sender based on his/her language typicality. In this context the aforementioned linguistic bound would not impose any limitation.

Our approach has been tested considering every message as an independent message providing heretofore certain confidence on our assumptions about the underlying essential behavioral patterns due to the obtained accuracy. However, this stringent policy is itself a detriment to our hypothesis, since most of such dynamic dialogues are short to be mined towards extracting fruitful properties [14] as preceding work has demonstrated in the past working with block sizes ranging from approximately 2000 words [12,13] to a minimum of 200 words [6]. A practical yet even more realistic workaround to the above eventuality can be straightforwardly implemented by voting the results of several consecutive model outcomes. In this manner the impersonation detection system would not focus on accurately classifying a single, hopefully discriminable message, but rather a series of texts in a much more similar fashion to chat sessions and messaging tools. In order to illustrate the performance of this scheme we have included results from this proposed voting approach as a function of the number of consecutive classification outcomes voted by majority. As shown in Table 4, voting over the decisions taken for successive messages enhances significantly the overall precision score. This strategy paves the way towards concatenating different messages before the essence extraction algorithm so as to better discriminate linguistically the set of possible authors. However, as long as more and more short messages are accumulated and jointly processed, the application scope of this work would shift towards more traditional long-text settings for authorship attribution, which falls out of the short-length messaging context within which this research work is framed.

6 Conclusions

Authorship attribution is conceived by the research community as the problem of identifying the origin of a text among different authors by solely analyzing

its content. This paradigm and the interesting technical approaches to efficiently solve it have embodied a very active research area so far, with a sharp multidisciplinary flavor due to the convergence of techniques and methods from Computational Intelligence, Machine Learning and Natural Language Processing. This paradigm has been mostly addressed from a literacy perspective, aiming at identifying the stylometric features and writeprints which unequivocally typify the writer patterns and allow their unique identification.

In this context, this article has hypothesized and analyzed the identification of the sender of a message as a useful approach to detect impersonation attacks in interactive communication scenarios. In particular conventional yet innovative characteristics of messages have been extracted via NLP techniques and selected by means of a newly devised feature selection algorithm based on the dissociation between essential traits of the sender and receiver influences. The proposed selection method has been shown to be promising with real SMS data in terms of identification accuracy, performance further enhanced by means of a more elaborated voting scheme using 1) soft estimates of the one-versus-one classifiers underlying beneath the overall authorship detection scheme; and 2) voting along estimates of different messages corresponding to a single communication session, as could be applied to e.g. chat sessions and message series.

With regard to future research lines derived from this work, it is important to note that the feature essence isolation approach proposed in this paper also provides information about the circumstances or conditions in which the dyadic discussion is framed. Thus, the discrimination of commonly used, essential features from conversations can shed light on the context of the communication scenario. Let us imagine an individual maintaining several dyadic communications. After discarding all the inherent and unconscious behavioral patterns, one can discover the different roles the user is deliberately developing, with very avant-garde applications such as the detection of pedophiles [34]. This future research direction paves the way to context clustering or audience categorization, with interesting extensions towards quantifying the linguistic essence of messages exchanged over a mixture of dyadic and non-dyadic communication channels. Extending the concept of linguistic essence to one-to-many conversational scenarios (as those held in public interfaces of Social Networks) will be also investigated in detail.

Acknowledgments

This work has been partially supported by the Basque Government under the ETORTEK (grant IE14-382) and the ELKARTEK (BID3A project, grant ref. KK-2015/0000080) funding programs.

References

[1] S. F. Hussain, A. Suryani, "On Retrieving Intelligently Plagiarized Documents using Semantic Similarity", Engineering Applications of Artificial Intelligence, Vol. 45, pp. 246-258, 2015.

[2] B. Galitsky, "Machine Learning of Syntactic Parse Trees for Search and Classification of Text", Engineering Applications of Artificial Intelligence, Vol. 26, N. 3, pp. 1072-1091, 2013.

[3] L. van der Knaap, F. Grootjen, "Author Identification in Chat Logs using Formal Concept Analysis", 19th Belgian-Dutch Conference on Artificial Intelligence, pp. 181–188, 2007.

[4] G. Inches, M. Harvey, F. Crestani, "Finding Participants in a Chat: Authorship Attribution for Conversational Documents", International Conference on Social Computing (SocialCom), pp. 272–279, 2013.

[5] M. L. Brocardo, I. Traore, S. Saad, I. Woungang, "Authorship Verification for Short Messages using Stylometry", IEEE International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1–6, 2013.

[6] G. Hirst, O. G. Feiguina, "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts", Literary and Linguistic Computing, Vol. 22, N. 4, pp. 405–417, 2007.

[7] N. Graham, G. Hirst, B. Marthi, "Segmenting Documents by Stylistic Character", Natural Language Engineering, Vol. 11, N. 4, pp. 397–415, 2005.

[8] A. Abbasi, H. Chen, "Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace", ACM Transactions on Information Systems, Vol. 26, N. 2, Article 7, 2008.

[9] D. Jurafsky, J. H. Martin, "Speech & Language Processing", Pearson Education India, 2009.

[10] F. J. Tweedie, R. H. Baayen, "How Variable may a Constant be? Measures of Lexical Richness in Perspective", Computers and the Humanities, Vol. 32, N. 5, pp. 323-352, 1998.

[11] E. Stamatatos, N. Fakotakis, G. Kokkinakis, "Automatic Text Categorization in Terms of Genre and Author", Computational Linguistics, Vol. 26, N. 4, pp. 471-495, 2000.

[12] H. Baayen, H. Van Halteren, F. Tweedie, "Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution", Literary and Linguistic Computing, Vol. 11, N. 3, pp. 121–132, 1996.

[13] P. M. McCarthy, G. A. Lewis, D. F. Dufty, D. S. McNamara, "Analyzing Writing Styles with Coh-Metrix", FLAIRS Conference, pp. 764-769, 2006.

[14] L. Pearl, M. Steyvers, "Detecting Authorship Deception: a Supervised Machine Learning Approach using Author Writeprints", Literary and Linguistic Computing, Vol. 27, N. 2, pp. 183–196, 2012.

[15] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods", Journal of the American Society for information Science and Technology, Vol. 60, N. 3, pp. 538-556, 2009.

[16] M. Fissette, F. A. Grootjen, "Author Identification in Short Texts", B.Sc. Thesis, Raboud Universiteit Nijmegen, 2013.

[17] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, "COMPA: Detecting Compromised Accounts on Social Networks", Symposium on Network and Distributed System Security (NDSS), 2013.

[18] E. Villar-Rodríguez, J. Del Ser, S. Salcedo-Sanz, "On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks", Springer Studies in Computational Intelligence, Vol. 570, pp. 259–268, 2015.

[19] K. Toutanova, D. Klein, C. D. Manning, Y. Singer, "Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network", Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol. 1, pp. 173–180, Association for Computational Linguistics, 2003.

[20] L. Derczynski, A. Ritter, S. Clarke, K. Bontcheva, "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data", International Conference on Recent Advances in Natural Language Processing, pp. pp. 198–206, 2013.

[21] I. S. Dhillon, S. Mallela, R. Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification", Journal of Machine Learning Research, Vol. 3, pp. 1265–1287, 2003.

[22] G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research, Vol. 3, pp. 1289–1305, 2003.

[23] M. Koppel, N. Akiva, I. Dagan, "Feature Instability as a Criterion for Selecting Potential Style Markers", Journal of the American Society for Information Science and Technology, Vol. 57, N. 11, pp. 1519-1525, 2006.

[24] J. Karlgren, G. Eriksson, "Authors, Genre, and Linguistic Convention", Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, 2007.

[25] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, B. Maia, "'twazn me!!!;(' Automatic Authorship Analysis of Micro-blogging Messages", In Natural Language Processing and Information Systems, pp. 161–168, 2011.

[26] T. Chen, M. Y. Kan, "Creating a Live, Public Short Message Service Corpus: the NUS SMS Corpus", Language Resources and Evaluation, Vol. 47, N. 2, pp. 299-335, 2013.

[27] T. A. Almeida, J. M. G. Hidalgo, A. Yamakami, "Contributions to the Study of SMS Spam Filtering: New Collection and Results", Proceedings of the 11th ACM symposium on Document Engineering, pp. 259–262, 2011.

[28] J. M. G. Hidalgo, T. A. Almeida, A. Yamakami, "On the Validity of a New SMS Spam Collection", IEEE International Conference on Machine Learning and Applications (ICMLA), Vol. 2, pp. 240-245, 2012.

[29] T. Almeida, J. M. G. Hidalgo, T. P. Silva, "Towards SMS Spam Filtering: Results under a new Dataset", International Journal of Information Security Science, Vol. 2, N. 1, pp. 1–18, 2013.

[30] L. Breiman, "Random Forests", Machine Learning, Vol. 45, N. 1, pp. 5–32, 2001.

[31] C. Cortes, V. Vapnik, "Support-Vector Networks", Machine learning, Vol. 20, N. 3, pp. 273-297, 1995.

[32] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods", Advances in Large Margin Classifiers, Vol. 10 (3), pp. 61–74, 2000.

[33] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", International Conference on Machine Learning (ICML), pp. 856–863, 2003.

[34] Mirror News: a real impersonation attack in Facebook, `http://www.mirror.co.uk/news/uk-news/facebook-paedophile-posed-teen-attempt-5563126`, published on April 22nd, 2015.

List of Tables

List of Figures

Table 1
Precision score for different supervised learning techniques and authorship classification approaches. Scores are given as *mean/standard deviation* computed over 10 stratified folds.

Approach (feature selection)	SVM	RF
Approach A (variance threshold, 19095 features)	0.718 / 0.001	0.607 / 0.014
Approach A (χ^2, best 300 features)	0.665 / 0.007	0.594 / 0.011
Approach A (χ^2, best 600 features)	0.679 / 0.011	0.609 / 0.014
Approach A (χ^2, best 900 features)	0.693 / 0.008	0.607 / 0.015
Approach A (ANOVA F-score, best 300 features)	0.668 / 0.009	0.600 / 0.018
Approach A (ANOVA F-score, best 600 features)	0.682 / 0.013	0.611 / 0.011
Approach A (ANOVA F-score, best 900 features)	0.688 / 0.011	0.611 / 0.017
Approach A (tree-based importance thresholding)	0.675 / 0.014	0.602 / 0.012
Approach A (Recursive Feature Elimination)	0.688 / 0.014	0.598 / 0.010
Approach A (Fast Correlation-Based Filter [33])	0.633 / 0.020	0.517 / 0.016
Approach A (PCA, 300 features)	0.687 / 0.011	0.412 / 0.013
Approach A (PCA, 600 features)	0.693 / 0.009	0.386 / 0.012
Approach A (PCA, 900 features)	0.692 / 0.006	0.384 / 0.013
Approach B (hard, $\Psi_{i,j} = 0.8$)	0.684 / 0.003	0.624 / 0.004
Approach B (hard, $\Psi_{i,j} = 0.6$)	0.649 / 0.001	0.599 / 0.002
Approach B (hard, $\Psi_{i,j} = 0.3$)	0.587 / 0.006	0.557 / 0.003
Approach B (hard, self-adjusted $\Psi_{i,j}$)	0.665 / 0.004	0.610 / 0.002
Approach B (soft, $\Psi_{i,j} = 0.8$)	0.710 / 0.003	0.629 / 0.002
Approach B (soft, self-adjusted $\Psi_{i,j}$)	0.703 / 0.004	0.618 / 0.002

Table 2
Absolute and relative number of features used for each approach (A & B) and threshold selection method.

Author	Approach A	Approach B			
		$\Psi_{i,j} = 0.8$	$\Psi_{i,j} = 0.6$	$\Psi_{i,j} = 0.3$	Self-adjusted $\Psi_{i,j}$
0	19095 (100%)	418 (2.19%)	156 (0.81%)	30 (0.15%)	215 (1.12%)
1		223 (1.17%)	92 (0.48%)	20 (0.10%)	157 (0.82%)
2		318 (1.66%)	147 (0.77%)	30 (0.16%)	184 (0.96%)
3		749 (3.92%)	305 (1.59%)	63 (0.33%)	288 (1.51%)
4		241 (1.26%)	107 (0.56%)	29 (0.15%)	135 (0.71%)
5		275 (1.44%)	129 (0.67%)	34 (0.18%)	161 (0.84%)
# unique/total features		920/2224	379/936	76/206	420/1140

Table 3
Normalized confusion matrix corresponding to Approach B, soft voting, self-adjusted $\Psi_{i,j}$.

		Predicted label					
		0	1	2	3	4	5
True label	0	0.68	0.08	0.02	0.06	0.15	0.01
	1	0.02	0.76	0.04	0.09	0.08	0.01
	2	0.04	0.11	0.65	0.08	0.11	0.01
	3	0.00	0.04	0.03	0.88	0.05	0.00
	4	0.03	0.04	0.06	0.05	0.33	0.48
	5	0.01	0.02	0.05	0.02	0.87	0.03

Table 4
Precision score (mean /std) for majority voting of successive message and its comparison to the figures of merit of Approaches A and B.

Approach	SVM	RF
A (no feature selection)	0.718 / 0.001	0.607 / 0.014
B (hard, self-adjusted $\Psi_{i,j}$)	0.665 / 0.004	0.610 / 0.002
B (soft, self-adjusted $\Psi_{i,j}$)	0.703 / 0.004	0.618 / 0.002
B (soft, self-adjusted $\Psi_{i,j}$, 3-msg. voting)	0.753 / 0.003	0.667 / 0.005
B (soft, self-adjusted $\Psi_{i,j}$, 5-msg. voting)	0.801 / 0.001	0.715 / 0.003
B (soft, self-adjusted $\Psi_{i,j}$, 7-msg. voting)	0.832 / 0.006	0.747 / 0.001
B (soft, self-adjusted $\Psi_{i,j}$, 9-msg. voting)	0.845 / 0.005	0.788 / 0.002

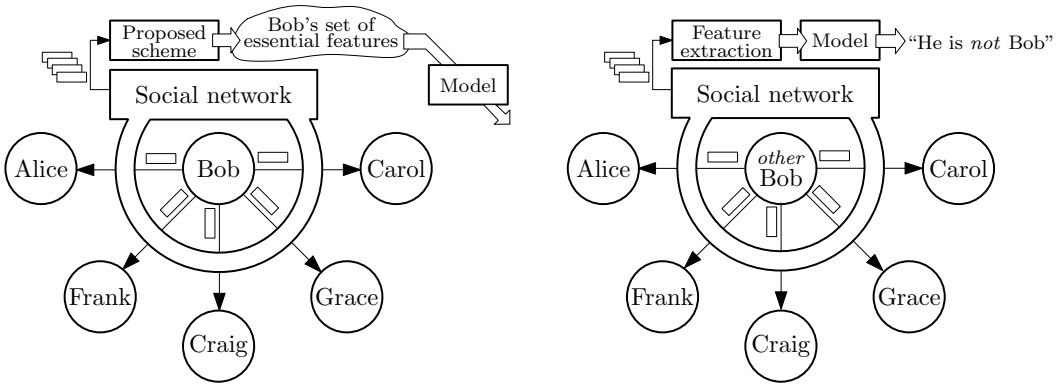

Fig. 1. Example of impersonation in social networks where the proposed feature selection scheme could serve to detect impersonated uses of any user's account.

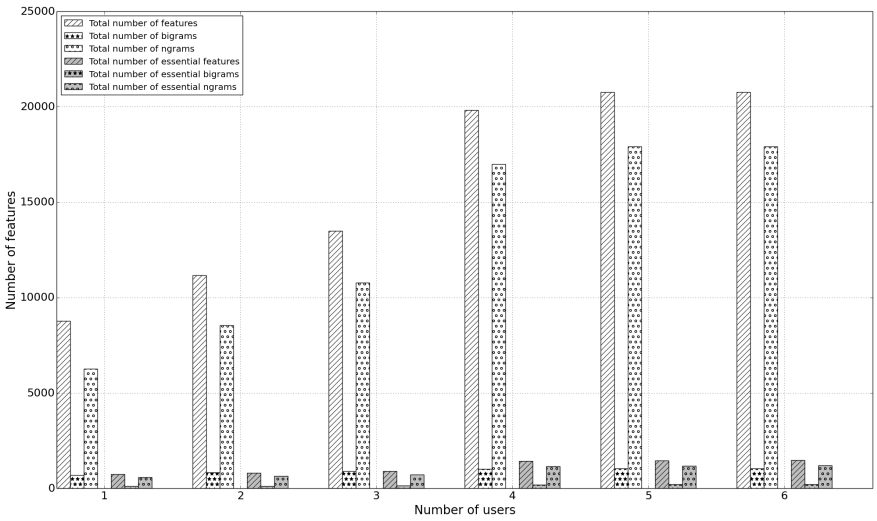

Fig. 2. Feature growth rate for both total and essential features versus the number of senders to be identified.

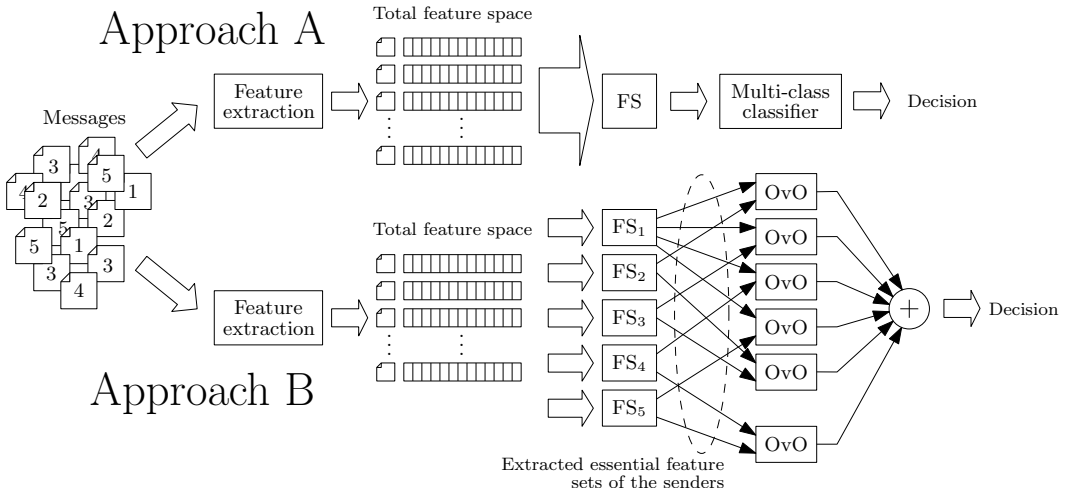

Fig. 3. Diagram showing the two considered sender identification approaches.

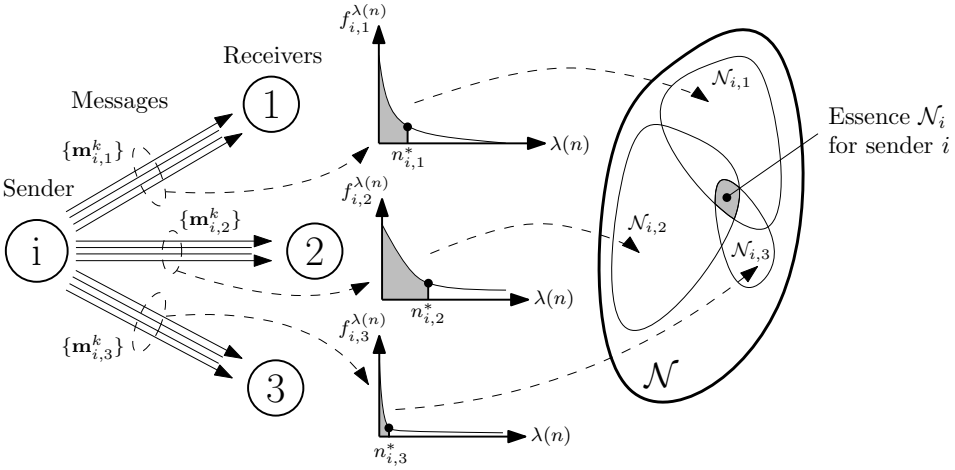

Fig. 4. Schematic diagram showing the essence extraction procedure for sender i.

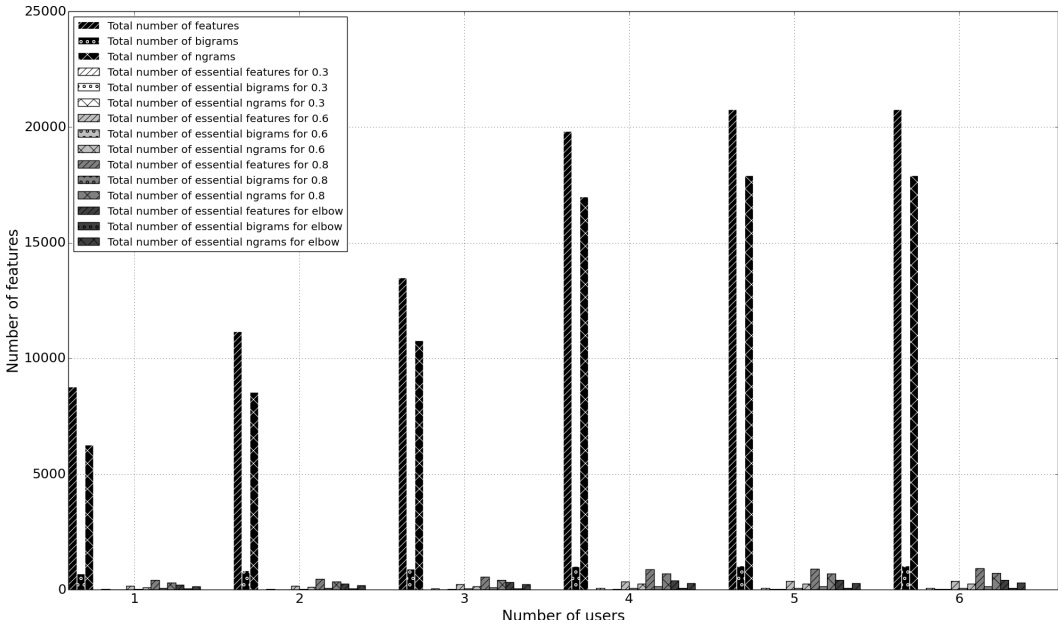

Fig. 5. Progression of the accumulated number of features per type (trigram, PoS bigram, essential) for different essence selection threshold schemes and number of considered users. It can be seen that the overall number of features does not increase when considering the last user (user 5), fact that unveils that a single message may not be sufficient for uniquely discriminating among certain authors, especially when dealing with datasets containing messages of reduced length.