Large-Scale Micro-Blog Authorship Attribution: Beyond Simple Feature Engineering*

Thiago Cavalcante, Anderson Rocha, and Ariadne Carvalho

Institute of Computing, University of Campinas
Av. Albert Einstein, 1251, Cidade Universitaria, Campinas/SP - Brasil, CEP 13083-852
thicoc@gmail.com, {anderson.rocha,ariadne}@ic.unicamp.br

Abstract. With the ever-growing use of social media, authorship attribution plays an important role in avoiding cybercrime, and helping the analysis of online trails left behind by cyber pranks, stalkers, bullies, identity thieves and alike. In this paper, we propose a method for authorship attribution in micro-blogs with efficiency one hundred to a thousand times faster than state-of-the-art counterparts. The method relies on a powerful and scalable feature representation approach taking advantage of user patterns in micro-blog messages, and also on a custom-tailored pattern classifier adapted to deal with big data and high-dimensional data. Finally, we discuss search-space reduction when analyzing hundreds of online suspects and millions of online micro messages, which makes this approach invaluable for digital forensics and law enforcement.

Keywords: Authorship Attribution, Micro-Blogs, Big Data.

1 Introduction

The recent explosion of social media brings about a great deal of freedom of speech, but this comes often along with anonymity. The problem is even more complex with deep web, the World Wide Web content that is not indexed by search engines. With so much freedom and possibilities of social connections, it is inevitable people start using anonymity as a weapon for their personal agenda, such as defaming politicians with opposite views, or impersonating one another. Recent media articles abound discussing online harassment cases, identity theft, online impersonation, stalking and alike to name just a few [14, 17]. A recent survey from the Wall Street Journal in the United States revealed that in 2011, more than 5% of Facebook, 6.3% of Twitter, and 7% of Google+ users suffered from identity theft [21]. The most alarming reality is that these numbers were up 13% with respect to the previous year.

The possibility of anonymity raises the need to identify authors by other means. The user messages themselves can be used for the task. Notwithstanding when it comes to social media, we have to make sense of a massive amount

* The authors thank the financial support of CAPES (Grant #01P45543013), CNPq (Grants #477662/2013-7, and #304352/2012-8), FAPESP (Grant #2010/05647-4), and Microsoft Research.

E. Bayro-Corrochano and E. Hancock (Eds.): CIARP 2014, LNCS 8827, pp. 399–407, 2014.

of data, transform it into information and, ultimately, into knowledge. Sophisticated learning solutions are needed for gleaning knowledge and insights from such data. To complicate matters even more, each message is, by itself, too small to allow the exploration of patterns and user trails. In this paper, we analyze micro-blog messages, more specifically Twitter data, whereby 500+ millions of new messages are exchanged everyday [8]. Easily, a cybercrime committed either by means of Twitter messages or commited elsewhere but with evidences on it, might have a large number of suspects, generating an even bigger amount of data to be analyzed and a real challenge for authorship attribution.

First, we approach the problem proposing a characterization technique that captures character and word properties used (character n-grams and word n-grams). Normally, these features were not explored in their full extent before in this context because they lead to high-dimensional feature vectors. Such vectors represent a real challenge for traditional classifiers to carve decision surfaces without being doomed by the curse of dimensionality. Differently, here we address this problem by showing an alternative for the traditional Support Vector Machine (SVM) classifier, which better deals with big data in terms of accuracy and performance. We rely upon Power Mean SVM [22], a solution recently proposed for large-scale visual classification and herein instantiated with success for the problem of large-scale authorship attribution of micro-blog messages. Finally, we discuss the viability of reducing the search space for forensic analysis when dealing with hundreds of suspects and millions of messages.

2 Related Work

Authorship attribution is a field of study that started outside computing. Many essays and even books were not published with the authors' real names, making it hard to identify the real authorship. A famous example is the one involving federalist papers, a collection of essays promoting the ratification of the US Constitution, some of them claimed both by Hamilton and Madison [12].

The first computational methods to infer the author of a text used a naïve-bayesian model [12]. This approach attempts to maximixe $P(\boldsymbol{x}|a)$ for a text $\boldsymbol{x}$ belonging to a candidate author a. Since then many authors have used the naïve-bayes model for the authorship attribution problem [10,13,15]. *Stylometry* was introduced to capture an author's style in terms of computational features.

A wide variety of the stylometric characteristics were suggested in the literature, including the use of text metrics such as average word and sentence lengh [12], and the use of n-grams. N-grams are the most widely used features, mainly because they can be applied to any sequence of tokens in the text such as characters, words, or even the semantic classes of words [7,19].

Since the bag-of-words model was successfully introduced, with the use of n-grams, another approach emmerged. In this approach, data can be represented in a vectorized form and used to train a general classifier. Among the classifiers used for authorship attribution are SVMs [3,4,15], decision trees [20,23], neural networks [23] and genetic algorithms [6].

Nowadays, we face a new challenge for authorship attribution due to the growth of social media, specifically with micro-blogs. They are composed of very short messages, which makes the analysis substantially harder. Although some authors have already worked with short mesages [5,15], only a few have tackled the authorship attribution problem in micro-blogs thus far [2,9,11,16,18].

Current literature shows that the methods relying upon the SVM classifier [11,16,18] outperform other approaches for authorship attribution on micro-blogs [16], namely naïve-bayes [2] and SCAP [9]. Almost all approaches use the same set of features, character-level and word-level n-grams [16]. These features normally follow standard practice on web data: $n = 4$ for character n-grams, and $n = (2, \ldots, 5)$ for word n-grams in a traditional bag-of-words model. This limitation was up to now mostly regarded to the impossibility of carving useful decision spaces with traditional implementations of the used classifiers.

3 Methodology

In this work, we propose a solution for authorship attribution on micro-blog texts. The innovative aspect of our approach is the use of a complete set of features relying on patterns extracted from unigrams to 5-grams of words.

Previous work did not use unigrams as they are allegedly captured by the character n-grams and due to the explosion in the feature representation harming the classification process. Differently, we show that unigrams, with the character n-grams, substantially improve the classification accuracy. We deal with the high-dimensional data representation using an improved version of the SVM classifier previously presented for large-scale image classification [22]. SVM LibLinear[1] does not solve this problem although it is faster than normal SVM.

In the following sections we describe our feature engineering for micro-blog attribution problem as well as the used classification technique.

Character N-grams. Character n-grams are often used for authorship attribution on web-based texts as they capture unusual features, such as emoticons and special use of punctuation. They help mending the effect of small typos authors do not repeat very often. For example, with the word "misspeling", the generated character 4-grams would still have "miss", "issp", "sspe", "spel" and "ling", in common with the 4-grams generated for the correct word "misspelling".

Following the literature [9,15,16], herein we focus on character 4-grams as whitespaces and metatags are included in the n-grams. Whitespaces are appended at the beginning and at the end of each tweet. Also we discard any character 4-grams which does not appear at least twice for the same author [16]. The features used are case-sensitive, since the author's preference for capitalization of letters is also one of her traits that can be used for identification.

Word N-grams. The use of the traditional bag-of-words model is proven to be useful for authorship attribution of micro-messages [16]. Because tweets are

[1] `www.csie.ntu.edu.tw/~cjlin/liblinear/`

very small, it is commonplace for users to repeat short phrases and the same words all over their micro-messages. We also include punctuation sequences to find the n-grams, considering that these sequences might be part of the phrases.

We used n-grams for words with $n \in (1, \ldots, 5)$. According to the literature, character level n-grams should generally capture the unigrams (1-grams) and their use would increase the feature vectors substantially. Herein we show their complementarity regarding the normal characterization used in the literature. Special meta-tags were also considered at the beginning and at the end of each tweet to distinguish words frequently used to start and to end messages. These features are also case-sensitive.

When using all n-grams and $n = 4$ word n-grams, the proposed method has feature vectors varying according to the number of training examples and analyzed users. For instance, the vector dimensionality varies from 20,000-d vectors (50 users and 50 training tweets per user) to around 500,000-d vectors (500 users and 500 tweets per user). Although the traditional literature for author attribution has used $n < 4$ as default (e.g., the authors in [4] used $n = 2$), for micro messages, we need a larger n to capture internet meta-language properties such as emoticons, onomatopoeia, abbreviations, and others.

Large-scale and High-dimensional Data Classification. While most authors use a traditional formulation of the SVM classifier, we observed it does not handle large data and high-dimensional feature vectors very well. This is indeed one of the reasons the literature has avoided unigrams for feature representation.

Differently, we rely upon Power Mean SVM kernel (PMSVM) formulation [22], which was originally proposed for large-scale image classification. The power mean kernel generalizes many kernels in the additive kernel family. These naturally arise in applications such as image and text classification, where the data is well represented by histograms or bag-of-word models. Also, this kernel family is not very sensitive to parametrization, avoiding overfitting the training data.

In a nutshell, the power mean kernel aggregates the advantages of linear SVM and non-linear additive kernel SVM. It performs faster than other additive kernels because instead of approximating the kernel function and the feature mapping, it approximates the gradient function using polynomial regression. This approach outperforms fast linear SVM solvers (e.g., LibLinear SVM, and Coordinate Descent SVM) in about 5×, and also, in the state-of-the-art additive kernel SVM training methods in about 2× (e.g., HIK SVM) [22]. Therefore, this kernel converges using only a small fraction of the iterations needed for the linear classification when dealing with a large number of features and big data [22].

This kernel is specially attractive for the nature of micro-blog messages; it provides very fast training times for the amount of data, and also will not overfit the sparse feature vectors generated in the bag-of-words model. For more details on this implementation, please refer to [22].

4 Experiments and Validation

Given that recent changes in the Twitter data policy, which forbids data exchange, we could not test our approach directly on the same dataset used by other authors. Therefore, to make a fair comparison between our methods and the state-of-the-art, we reproduced the experiment as described in Koppel et al. [16] and implemented their best method. The authors proposed the current state-of-the-art method, and performed an extensive comparison with other works for micro-blog authorship attribution [2, 9], outperforming them all.

For each test, we performed 10 runs. The users in each run were choosen at random from the dataset. For each run, we followed a 10-fold cross-validation protocol reporting the average classification accuracy.

Dataset and Pre-Processing. The dataset comprises 10^7 tweets from 10^4 writers in English. Each tweet is at most 140-character long and may include hashtags, user indications and links. The dataset was collected across several days with the latest 3,200 tweets from each user. To restrict the search to English speaking users, we searched for English function words in the API [9].

The pre-processing of each micro-message includes removing all non-english tweets, tweets with less than three words and retweets, those marked as retweets or any tweet containing the meta-tag RT. For sparsity reasons, we replaced numbers, URLs, dates and timestamps by the metatags NUM, URL, DAT, TIM. Moreover, the hashtags and user references were replaced, since they enrich the feature set for author attribution [9].

Unigrams as Features and Comparison with the Literature. Although unigrams have not been used previously for micro-blogs authorship attribution for the reasons discussed before, our findings show that they offer a substantial contribution to the problem. The use of unigrams generates a greater number of features to be used. Since we use a solution that relies upon PMSVM, the number of features has less impact on the classification time, improving accuracy significantly.

Using unigrams for feature representation, our solution outperforms the literature by a margin of over 10% for 50 users and 500 tweets per user, as shown in Fig. 2. Most importantly, the method proposed in the literature did not converge for all tests in more than two days of computing when we used 1,000 training tweets per user. This is a significant margin explained by the use of a classifier adequate for big data and a large number of features. Note the steady improvement in the classification task allowed by the use of unigrams as features.

We also evaluated how our approach deals with an increasing number of users, testing our solution with 500 users. Fig. 4 shows the results for these experiments. We can see that that the classifier nicely handles hundreds of users and continues to perfect as more training messages are used per user.

Power Mean SVM Analysis. Here we explore how the proposed method compares with others with respect to the computational efficiency considering 50 users and a varying number of training messages per user. Fig. 1 shows

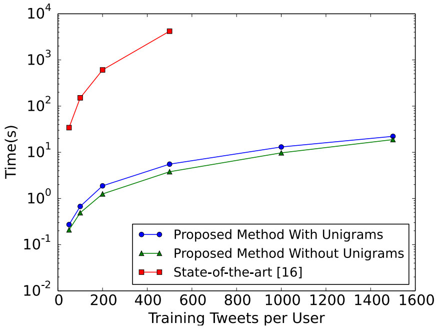

Fig. 1. Efficiency comparison between traditional SVM-based solutions and ours based on PMSVM for 50 users

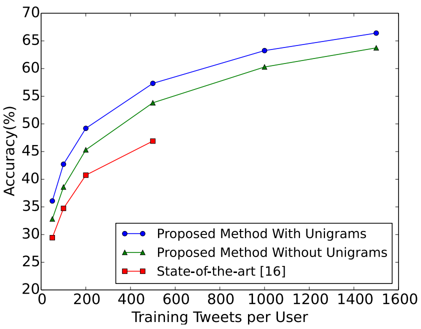

Fig. 2. Our approaches *v.* the state-of-the-art [16] for 50 users

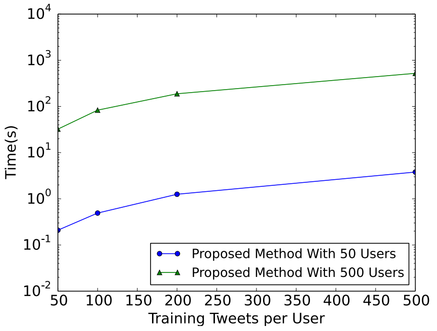

Fig. 3. Time consumption of the proposed method for 50 and 500 users

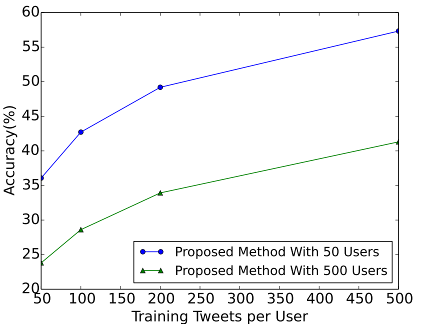

Fig. 4. Accuracy of the proposed method for 50 and 500 users

that the proposed classification technique runs between 100× to 1,000× faster than [16], which relies on traditional SVM formulation. Our solution based on PMSVM also outperforms [16] in terms of classification effectiveness. Using 50 tweets per user for training the method is 2.5% more effective; when using 500 training tweets per user, the difference between the methods increases to more than 4% as Fig. 2 depicts. This difference tends to increase as more training data is used as the proposed solution nicely handles large-scale data. Note again that PMSVM is neither sensitive to the parameter C nor other parameters of the SVM classifier, so there is no need for a parameter grid search for fine tuning [22].

For this analysis, we use a 2.30GHz 3rd generation Intel Core i7-3610QM computer with 8GB 1600MHz DDR3 SDRAM (2 DIMM) running a Fedora Linux. For better visualisation, we used a log scale. Both curves approach a quadratic function of the number of training data.

Search-Space Reduction. Although traditional authorship analysis intends to find the author of a given text, that might not be the case with micro web messages. The very nature of short messages make it hard to attribute them to a single author. Also, the problems involved in the task, such as hoaxes, impersonations and identity stealing, make it an open problem where the user may not be among the suspects, and the classifiers will always point out to someone. This suggests that the effort should also be on reducing and prioritizing the suspects rather than pointing to a single culprit.

Instead of seeking for the most probable user, we can rank the classes according to the output function, and then show how well we can reduce the search space of the problem. We tested the method with 500 users and a varying number of tweets per user. The Cumulative Matching Curve (CMC) shows the accuracy of finding the author of a tweet considering the top N users.

The classifier starts with a classification accuracy of 35% when using 50 tweets per user for training. Considering the random baseline of 0.2%, this is a remarkable result for micro-blog authorship attribution. In more than 65% of the cases, the correct user will be among the top 50 users when we use 200 training tweets per user (see Fig. 5). In a real scenario, this would reduce the number of suspects to 10% of the original size in more than half of the scenarios.

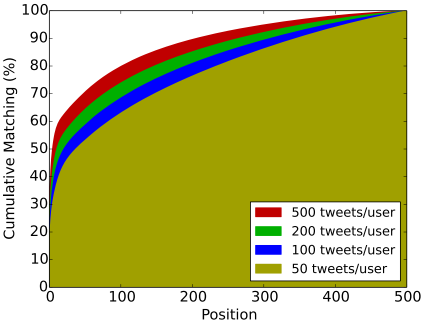

Fig. 5. CMC for 500 users with a varying number of training tweets per user

5 Conclusions and Future Work

This article showed that the information explosion and scarcity of information per user in micro-blog social networks require better methods to cope with the problem. Even when dealing with micro-messages, the problem grows really fast, because of the large number of users and messages involved. Also, there is a huge amount of features to be considered, so as to make them really representative and unique, specially due to the use of unconventional punctuation, abbreviations and internet meta-language. This generates feature vectors with high dimensions, which in turn can lead to the curse of dimensionality [1]. Therefore, we need to use classifiers that better handle large datasets, as well as the high number of dimensions generated by the problem.

Although the accuracy of the classification does not hit perfection, the method discussed herein is surely an advance when compared to the literature and opens the possibility of other authors to explore features and enhanced classifiers overseen thus far. In addition, the cumulative matching approach analysis shows that our method can greatly reduce the number of users to be analyzed in a real situation. We also showed that some features, which capture great stylistic patterns, not used before due to technical limitations, now can be considered because of their discrimination power.

For future work, we plan to include the flexible pattern features suggested by Koppel et al. [16] to further improve accuracy. Also, we need to explore other non-textual features of social media, such as social graph.

References

1. Bishop, C.M.: Pattern recog. and machine learning, vol. 1. Springer (2006)
2. Boutwell, S.R.: Authorship attribution of short messages using multimodal features. Master's thesis, Naval Postgraduate School, Monterey, CA, USA (2011)
3. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Applied Intelligence 19(1-2), 109–123 (2003)
4. Forstall, C.W., Scheirer, W.J.: Features from frequency: Authorship and stylistic analysis using repetitive sound. In: Annual Chicago Colloquium on Digital Humanities and Computer Science (2009)
5. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing 22(4), 405–417 (2007)
6. Holmes, D.I., Forsyth, R.S.: The federalist revisited: New directions in authorship attribution. Literary and Linguistic Computing 10(2), 111–127 (1995)
7. Juola, P.: Authorship attribution. Foundations and Trends in information Retrieval 1(3), 233–334 (2006)
8. Krikorian, R.: New tweets per second record, and how! Twitter Blog (2013), `http://tinyurl.com/kcuhdcw` (accessed on May, 2014)
9. Layton, R., Watters, P., Dazeley, R.: Authorship attribution for twitter in 140 characters or less. In: Cybercrime and Trustworthy Computing, pp. 1–8 (2010)
10. Madigan, D., Genkin, A., Lewis, D.D., Lewis, E.G.D.D., Argamon, S., Fradkin, D., Ye, L., Consulting, D.D.L.: Author identification on the large scale. In: Meeting of the Classification Society of North America (2005)
11. Mikros, G.K., Perifanos, K.: Authorship attribution in greek tweets using authors multilevel n-gram profiles. In: AAAI Spring Symposium Series (2013)
12. Mosteller, F., Wallace, D.L.: Inference and Disputed Authorship: The Federalist Papers. Addison-Wesley, Reading (1964)
13. Peng, F., Schuurmans, D., Wang, S.: Augmenting naive bayes classifiers with statistical language models. Information Retrieval 7(3-4), 317–345 (2004)
14. Ramshaw, E.: Bashing the candidates with their own names. The New York Times (May 2012), `http://tinyurl.com/q6lc2fw` (accessed on May, 2014)
15. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In: Conference on Empirical Methods in Natural Language Processing, pp. 482–491. Association for Computational Linguistics (2006)
16. Schwartz, R., Tsur, O., Rappoport, A., Koppel, M.: Authorship attribution of micro-messages. In: Conference on Empirical Methods on Natural Language Processing, pp. 1880–1891. ACL (2013)

17. Shih, G.: Anonymous twitter feeds arise as political weapon. The New York Times (June 2014), http://tinyurl.com/5vol3gt (accessed on May, 2014)
18. Silva, R.S., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., Maia, B.: Twazn me!!!(automatic authorship analysis of micro-blogging messages. In: Natural Language Processing and Information Systems, pp. 161–168. Springer (2011)
19. Stamatatos, E.: A survey of modern authorship attribution methods. J. of the American Society for Information Science and Technology 60(3), 538–556 (2009)
20. Uzuner, Ö., Katz, B.: A comparative study of language models for book and author recognition. In: Intl. Joint Conf. on Natural Language Processing, p. 969 (2005)
21. Waters, J.: Why id thieves love social media. The Wall Street Journal (March 2012), http://tinyurl.com/ldvhpsb (accessed on May, 2014)
22. Wu, J.: Power mean svm for large scale visual classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2344–2351 (2012)
23. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. J. of the American Society for Information Science and Technology 57(3), 378–393 (2006)