'twazn me!!! ;(' Automatic Authorship Analysis of Micro-Blogging Messages

Rui Sousa Silva[1,3], Gustavo Laboreiro[2,4], Luís Sarmento[2,4], Tim Grant[1], Eugénio Oliveira[2], and Belinda Maia[3]

[1] Centre for Forensic Linguistics at Aston University
[2] Faculdade de Engenharia da Universidade do Porto - DEI - LIACC
[3] CLUP - Centro de Linguística da Universidade do Porto
[4] SAPO Labs Porto

Abstract. In this paper we propose a set of stylistic markers for automatically attributing authorship to micro-blogging messages. The proposed markers include highly personal and idiosyncratic editing options, such as 'emoticons', interjections, punctuation, abbreviations and other low-level features. We evaluate the ability of these features to help discriminate the authorship of Twitter messages among three authors. For that purpose, we train SVM classifiers to learn stylometric models for each author based on different combinations of the groups of stylistic features that we propose. Results show a relatively good-performance in attributing authorship of micro-blogging messages ($F = 0.63$) using this set of features, even when training the classifiers with as few as 60 examples from each author ($F = 0.54$). Additionally, we conclude that emoticons are the most discriminating features in these groups.

1 Introduction

In January 2010 the *New York Daily News* reported that a series of Twitter messages exchanged between two childhood friends led to one murdering the other. The set of Twitter messages exchanged between the victim and the accused was considered a potential key evidence in trial, but such evidence can be challenged if and when the alleged author *refutes* its authorship. *Authorship analysis* can, in this context, contribute to confirming or excluding the hypothesis that a given person is the true author of a queried message, *among several candidates*. However, the micro-blogging environment raises new, significant challenges as the messages are *extremely short* and fragmentary. For example, Twitter messages are limited to 140 characters, but very frequently have only 10 or even fewer words. Standard stylistic markers such as *lexical richness*, *frequency of function words*, or *syntactic measures* — which are known to perform well with longer, 'standard' language texts — perform worse with such short texts, whose

R. Muñoz et al. (Eds.): NLDB 2011, LNCS 6716, pp. 161–168, 2011.

language is 'fragmentary' [1]. Traditional authorship analysis methods are considered unreliable for text excerpts smaller than 250-500 words, as the accuracy tends to drop significantly with text length decrease [9].

In this paper we use a text classification approach to investigate whether some 'non-traditional' stylistic markers, such as the type of emoticons, provide enough stylistic information to be used in authorship attribution. We focus specifically on *Twitter* for its popularity, and address Portuguese in particular, which is one of the most widely used languages in this medium[1].

2 Related Work

In recent years, there has been considerable research on authorship attribution of some *user-generated contents* — such as *e-mail* (e.g. [2]) and, more recently, *web logs* (e.g. [3,4,5]) and 'opinion spam' (e.g. [6]). However, research on authorship attribution of Twitter messages has been scarce, and raised robustness problems.

To tackle the problem of robustness in computational stylometric analysis, research (e.g. the 'Writeprints technique' [10]) was applied to four different text genres to discriminate authorship and detect similarity of online texts among 100 authors. The performance obtained was good, but (a) the procedure did not prove to be content-agnostic, and (b) did not analyse Twitter messages. Also, using structural features that are possibly due to editing and considering 'idiosyncratic features' usage anomalies to include misspellings and grammar mistakes it is bound to compromise the results.

More recently, it has been demonstrated that the authorship of twitter messages can be attributed with a certain degree of certainty [11]. Surprisingly, the authors concluded that authorship could be identified at 120 tweets per user, and that more messages would not improve accuracy significantly. However, their method compromises the authorship identification task of most unknown messages, as they reported a loss of 27% accuracy when information about the interlocutor's user data was removed.

It has also been demonstrated that authorship could be attributed using 'probabilistic context-free grammars' [12] by building complete models of each author's (3 to 6) syntax. Nevertheless, the authors used both syntactic and lexical information to determine each author's writing style.

Conversely, we propose a content-agnostic method, based on low-level features to identify authorship of unknown messages. This method is independent of user information, so not knowing the communication participants is irrelevant to the identification task. Moreover, although some of the features used have been studied independently, this method is innovative in that the specific combination of the different stylistic features has never been used before and has not been applied to such short texts.

[1] `http://semiocast.com/downloads/Semiocast_Half_of_messages_on_Twitter_are_not_in_English_20100224.pdf`

3 Method Description and Stylistic Features

Authorship attribution can be seen as a typical *text classification* task: given examples of messages written by a set of authors (classes), we aim to attribute authorship of messages of unknown authorship. In a forensic scenario, the task consists of discriminating the authorship of messages of a small number of potential authors (e.g. 2 to 5), or determining whether a message can be attributed to a certain ('suspect') author.

The key to framing authorship attribution as a text classification problem is the selection of the feature sets that best describe the *style* of the authors. We propose four groups of stylistic features for automatic authorship analysis, each dealing with a particular aspect of tweets. All features are *content-agnostic*; to ensure a robust authorship attribution and prevent the analysis from relying on topic-related clues, they do not contain lexical information.

Group 1: Quantitative Markers. These features attempt to grasp simple quantitative style markers from the message as a whole. The set includes message statistics, e.g. length (in characters) and number of tokens, as well as token-related statistics (e.g. average length, number of 1-character tokens, 2-consonant tokens, numeral tokens, choice of case, etc). We also consider other markers, e.g. use of dates, and words not found in the dictionary[2] to indicate possible spelling mistakes or potential use of specialised language.

As *Twitter*-specific features, we compute the number of user references (e.g. @user_123), number and position of *hashtags* (e.g. #music), in-message URLs and the URL shortening service used. We also take note of messages starting with a username (a reply), as the author may alter their writing style when addressing another person.

Group 2: Marks of Emotion. Another highly personal — and hence idiosyncratic stylistic marker — is the device used to convey emotion. There are mainly three non-verbal ways of expressing emotion in user-generated contents: (i) *smileys*; (ii) *'LOLs'*; and (iii) *interjections*.

Smileys (':-)') are used creatively to reflect human emotions by changing the combination of eyes, nose and mouth. This work explores three axes of idiosyncratic variation: *range* (e.g. number of happy smileys per message), *structure* (e.g. whether the smiley has a nose) and *direction* of the smiley.

Another form of expression is the prevalent 'LOL', which usually stands for *Laughing Out Loud*. Frequently users manipulate the basic 'LOL' and 'maximise' it in various other forms, e.g. by repeating its letters (e.g. 'LLOOOLLL') or creating a loop (e.g. 'LOLOL'). This subgroup describes several instances of *length*, *case* and *ratio* between 'L' / 'O', so as to distinguish between 'LOL' and the exaggeration in multiplying the 'O', as in 'LOOOOL'.

We identify interjections as tokens consisting of only two alternating letters that are not a 'LOL', such as 'haaahahahah'. Other popular and characteristic

[2] We use the GNU Aspell dictionary for European Portuguese.

examples are the typical Brazilian laughing 'rsrsrs' and the Spanish laughing 'jejeje' — both of which are now commonly found in European Portuguese *Twitter*. We count the number of interjections used in a message, their average length and number of characters.

Group 3: Punctuation. The choice of punctuation is a case of writing style [13], mostly in languages whose syntax and morphology is highly flexible (such as Portuguese and Spanish). Some authors occasionally make use of expressive and non-standard punctuation, either by repeating ('!!!') or combining it ('?!?'). Others simply skip punctuation, assuming the meaning of the message will not be affected. Ellipsis in particular can be constructed in less usual ways (e.g. '..' or '.....'). We count the frequency of these and other peculiar cases, such as the use of punctuation after a 'LOL' and at the end of a message (while ignoring URLs and *hashtags*).

Group 4: Abbreviations. Some abbreviations are highly idiolectal, thus depending on personal choice. We monitor the use of three types of abbreviations: 2-consonant tokens (e.g. 'bk' for 'back'), 1- or 2-letter tokens followed by '.' or '/' (e.g. 'p/') and 3-letter tokens ending in two consonants, with (possibly) a dot at the end (e.g. 'etc.').

4 Experimental Setup

This study is focused on the authorship identification of a message among three candidate authors. We consider only three possible authors as forensic linguistic scenarios usually imply a limited number of suspect authors, and is hence more realistic. We chose to use Support Vector Machines (SVM) [14] as the classification algorithm for its proven effectiveness in text classification tasks and robustness in handling a large number of features. The *SVM-Light* implementation [14] has been used, parametrised to a linear kernel. We employ a *1-vs-all* classification strategy; for each author, we use a SVM to learn the corresponding stylistic model, capable of discriminating each author's messages. Given a suspect message from each author, we use each SVM to predict the degree of likelihood that each author is the true author. The message authorship is attributed to the author of the highest scoring SVM. We also consider a threshold on the minimum value of the SVM score, so as to introduce a *confidence* parameter (the minimum score of the SVM classifier considered valid) in the authorship attribution process. When none of the SVM scores achieves the minimum value, authorship is left undefined.

Our data set consists of *Twitter* messages from authors in Portugal, collected in 2010 (January 12 to October 1). We counted over 200,000 users and over 4 million messages during this period (excluding messages posted automatically, such as news feeds). From these, we selected the 120 most prolific *Twitter* authors in the set, responsible for at least 2,000 *distinct* and *original* messages (i.e. excluding *retweets*), to extract the sets of messages for our experiments. We

divide the 120 authors into 40 groups of 3 users at random, and maintain these groups throughout our experiments. The group of 3 authors forms the basic testing unit of our experiment.

We perform two sets of experiments. In *Experimental Set 1*, the classification procedure uses all possible groups of features to describe the messages. We use data sets of sizes 75, 250, 1,250 and 2,000 messages/author. In *Experimental Set 2*, we run the training and classification procedure using *only one* group of features at a time. We use the largest data set from the previous experiment (2,000 messages/author) for this analysis. We measure *Precision* (P), *Recall* (R) and F ($2PR/(P+R)$) considering:

$$P = \frac{\#\text{ messages correctly attributed}}{\#\text{ messages attributed}} \qquad R = \frac{\#\text{ messages correctly attributed}}{\#\text{ messages in the set}}$$

We run the training and classification procedures in each set of experiments and use the *confidence* parameter to draw *Precision vs. Recall graphs*. As these experiments consider *three* different authors, the baseline is $F = 0.33$ ($P = 0.33$ at $R = 0.33$). All experiments were conducted using a 5-fold cross validation, and run for all 40 groups of 3 authors. For varying levels of Recall (increments of 0.01) we calculate the maximum, minimum and average Precision that was obtained for all 40 groups. All F values are calculated using the average Precision.

5 Results and Analysis

Figure 1 shows the Precision vs Recall graphs for Experimental Set 1. Data set increases (from 75 to 2,000 messages/author) returns improvements in the

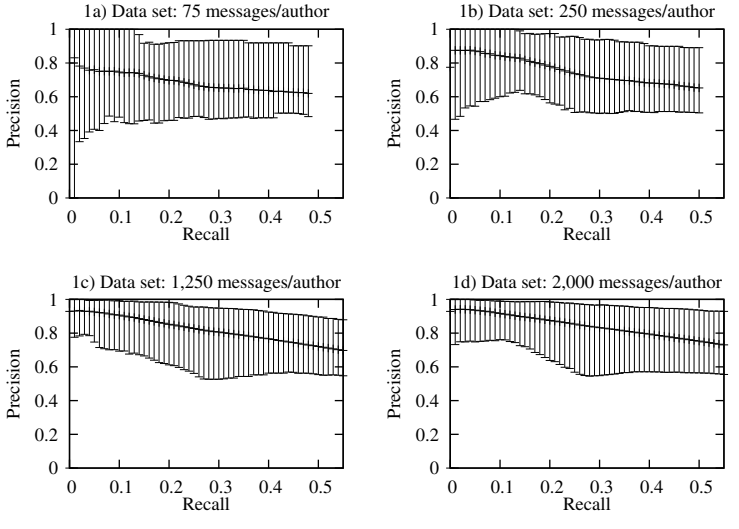

Fig. 1. Performance of each data set size. Each graph plots maximum, average and minimum Precision at varying levels of Recall (40 groups of 3 authors).

minimum, maximum and average Precision values. In addition, the robustness of the classifier also benefits from the added examples, as the most problem situations (corresponding to the minimum precision values) are handled correctly more frequently. The best F values are always obtained at the highest value of Recall, meaning that they too follow this improvement trend.

For the smaller data set (75 messages, Figure 1a), the minimum Precision curve is nearly constant, not showing a benefit from the decision threshold (at the cost of Recall). We speculate this is due to two reasons. First, given the large feature space (we use at least 5680 dimentions), and the relatively small number of non-negative feature component in each training example (most messages have between 64 and 70 features), a robust classification model can only be inferred using a larger training. Second, with such small sets it is highly probable that both the training and the test sets are atypical and distinct in terms of feature distribution. Still, the performance values obtained are far above the baseline, and an F value of 0.54 is reached. In the larger data sets (Figures 1c and 1d) we always obtain a Precision greater than 0.5. This means that even in the more difficult cases, the attribution process is correct more often than not. However, the contribution of the extra examples for the F values is lower when we go beyond 250 messages/author (where we get 0.59), even if they increase almost linearly up to 0.63 (for 2,000 messages/author).

Figure 2 presents the Performance vs Recall curves for authorship attribution with a classification procedure using *only one* group of features at a time. Quantitative Markers (Group 1) show an average performance, with minimum Precision

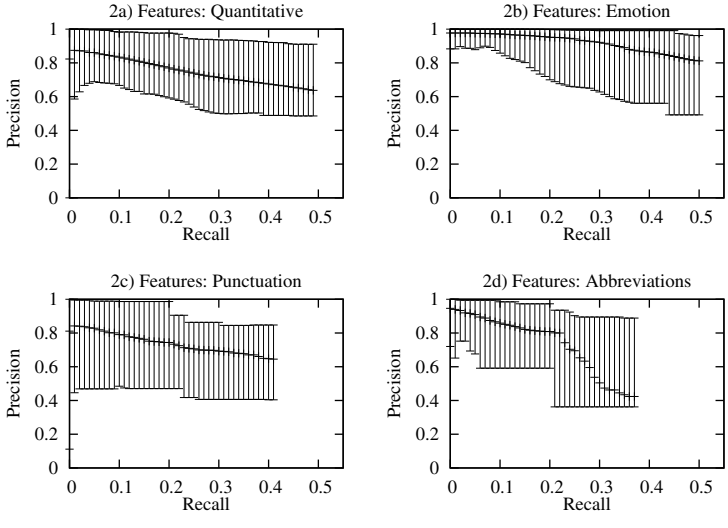

Fig. 2. Performance of each individual set of features. Each graph plots maximum, average and minimum Precision at varying levels of Recall (40 groups of 3 authors, 2,000 messages from each author).

and maximum Recall of 0.49, and maximum F value of 0.55 (Figure 2a). This shows that, albeit Twitter length constraints, there is room for stylistic choices like length of tokens, length of message posted, etc. Markers of *expression of emotion*, including *smileys*, *LOLs* and *interjections* (Group 2) achieve a relatively high performance, and clearly outperform all other feature groups (Figure 2b). It achieves an F value of 0.62 (where using *all* features togheter achieves 0.63). This is particularly interesting since these features are specific to user-generated contents, and to our knowledge their relevance and effectiveness in authorship attribution is now quantified for the first time. The difference between the average and minimum Precision values is an indicator that the low performance of this feature group is an infrequent event. The group of features including punctuation (Group 3) performs slightly worse than the previous groups, and scores only 0.50 on the F measure (Figure 2c). The difference between the best and worse case is significant, but the average Precision degrades as the Recall increases. Our evaluation demonstrates that our approach, although quite simplistic, is capable of detecting stylistic variation in the use of punctuation, and of successfully using this information for authorship attribution. This result is in line with those reported previously by [8] for punctuation-based features applied to automatic authorship attribution of sentences from newspapers. Group 4, containing features on the use of abbreviation, led to the worst results (maximum F value of 0.40). The shape of the curve rapidly approaches the baseline values, proving that this group is not robust (Figure 2d). Manual evaluation shows that these abbreviations are used rarely. However, as the low Recall/high Precision part of the curve suggests, they carry stylistic value, in spite of being used only in a relatively small number of cases. Finally, the performance when using *all* groups of features simultaneously (Figure 1f) is better than using any group of features *individually*, showing that all individual groups of features carry relevant stylistic information that can be combined, and suggesting that the investment in devising new groups of stylistic features may lead to additional global performance improvements — especially the recall.

6 Conclusions

Our experiment demonstrates that standard text classification techniques can be used in conjunction with a group of content-agnostic features to successfully attribute authorship of Twitter messages to three different authors. Automatic authorship attribution of such short text strings, using only *content-agnostic* stylistic features, had not been addressed before. Our classification approach requires a relatively small amount of training data (as little as 100 example messages) to achieve good performance in discriminating authorship.

Surprisingly, the group of emoticons outperforms all other feature groups tested, with a relatively high performance. The relevance and effectiveness of these features for automatic authorship attribution are now demonstrated for the first time. Quantitative and punctuation markers show average results, carrying some idiolectal information, despite the text length constraints. On balance, it can be argued that all features carry relevant information, since using

all groups of features simultaneously allows inferring more robust authorship classifiers than using any group of features individually.

Acknowledgments. This work was partially supported by grant SFRH/BD/47890/2008 FCT-Portugal, co-financed by POPH/FSE.

References

1. Grant, T.: Txt 4n6: Idiolect free authorship analysis. In: Coulthard, M., Johnson, A. (eds.) Routledge Handbook of Forensic Linguistics. Routledge, New York (2010)
2. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining e-mail content for author identification forensics, vol. 30, pp. 55–64. ACM, New York (2001)
3. Park, T., Li, J., Zhao, H., Chau, M.: Analyzing writing styles of bloggers with different opinions. In: Proceedings of the 19th Annual Workshop on Information Technologies and Systems (WITS 2009), Phoenix, Arizona, USA, December 14-15 (2009)
4. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers' age and gender. In: International AAAI Conference on Weblogs and Social Media (2009)
5. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology 60(1), 9–26 (2009)
6. Jindal, N., Liu, B.: Opinion spam and analysis. In: WSDM 2008: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 219–230. ACM, New York (2008)
7. Pavelac, D., Justino, E., Olivera, L.S.: Author identification using stylometric features. Intelligencia Artificial,Revista Iberoamericana de IA 11(36), 59–66 (2007)
8. Sousa-Silva, R., Sarmento, L., Grant, T., Oliveira, E.C., Maia, B.: Comparing sentence-level features for authorship analysis in portuguese. In: PROPOR, pp. 51–54 (2010)
9. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Lit. Linguist. Computing 22(4), 405–417 (2007)
10. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26(2), 1–29 (2008)
11. Layton, R., Watters, P., Dazeley, R.: Authorship attribution for twitter in 140 characters or less. In: Workshop Cybercrime and Trustworthy Computing, pp. 1–8 (2010)
12. Raghavan, S., Kovashka, A., Mooney, R.: Authorship attribution using probabilistic context-free grammars, pp. 38–42 (2010)
13. Eagleson, R.: Forensic analysis of personal written texts: a case study. In: Gibbons, J. (ed.) Forensic Linguistics: An Introduction to Language in the Justice System, pp. 362–373. Longman, Harlow (1994)
14. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)