Mining Online Diaries for Blogger Identification

Haytham Mohtasseb and Amr Ahmed *

Abstract—**In this paper, we present an investigation of authorship identification on personal blogs or diaries, which are different from other types of text such as essays, emails, or articles based on the text properties. The investigation utilizes couple of intuitive feature sets and studies various parameters that affect the identification performance.**

Many studies manipulated the problem of authorship identification in manually collected corpora, but only few utilized real data from existing blogs. The complexity of the language model in personal blogs is motivating to identify the correspondent author. The main contribution of this work is at least three folds. Firstly, we utilize the LIWC and MRC feature sets together, which have been developed with Psychology background, for the first time for authorship identification on personal blogs. Secondly, we analyze the effect of various parameters, and feature sets, on the identification performance. This includes the number of authors in the data corpus, the post size or the word count, and the number of posts for each author. Finally, we study applying authorship identification over a limited set of users that have a common personality attributes. This analysis is motivated by the lack of standard or solid recommendations in literature for such task, especially in the domain of personal blogs.

The results and evaluation show that the utilized features are compact while their performance is highly comparable with other larger feature sets. The analysis also confirmed the most effective parameters, their ranges in the data corpus, and the usefulness of the common users classifier in improving the performance, for the author identification task.

Keywords: Web Mining, Information Extraction, Psycholinguistic, Machine Learning, Authorship Identification

1 Introduction

Blog, or Web Log, is one of the most popular web media which allow people to write about their ideas and update the content in a chronological order. Recently the content of the web is rapidly changing, which opens new directions of use, allows collaboration all over the world, and collecting large amount of text. Previously, the web site owners have the control over the published materials. But now, web site users take up this role, at least partially. Users can create web pages, add photos and videos, write reviews, and express their feelings and emotions.

Blogs are one of the most popular forms of users' contribution to the web contents. There are many categorizations of blogs which are different in the content, publishing methodology, and even in the type of readers. Personal blog, or online diary, is the most famous category in which the blogger expresses his/her feelings, show creativity, and communicate with other people faster than emails or any other media. In addition, there are some targeted or focused blogs which focus on a specific subject such as news blogs, political blogs, and educational blogs. Our research is focused on the personal blogs category. We selected one of the famous personal blog sites, namely the "LiveJournal"[1]. LiveJournal is a free personal blog website forming a community on the internet that contains millions of users publishing their own ongoing personal diaries.

The availability of such text collections on the web has attracted the attention of researchers to apply text classification to induce the topic, opinion, mood, and personality. One of the active research areas in text classification is Authorship Identification. Authorship identification is the process of discovering or distinguishing the author of a given particular text from a set of candidate authors. Authorship identification is one of the authorship analysis tasks which include also similarity detection that evaluates the similarity between different text documents regardless of the authors of the text. The clear difference between the two types of authorship analysis is that the classes in authorship identification are predefined while there are no specified classes for similarity detection.

Authorship identification in blogs has various motivations and challenges. Identifying the author of anonymous blog posts could be useful in various applications. This includes online security where it is valuable to extract the patterns of authors who may participate in different blog sites with different identities. However, the task has its associated challenges. The large number of authors is one of the key factors in authorship identification. In particular, scaling existing solutions with the huge, and increasing, number of authors is a challenge. Moreover, there are many factors that have important roles and

*Department of Computing & Infomratincs, University of Lincoln, UK. Email: hmohtasseb@lincoln.ac.uk

[1]http://www.livejournal.com

affect the performance of identification process such as the text length, the number of posts per author, and the type of authors. There are many studies in this area, the authorship identification, on different types of text like emails, books, web forums, articles, and a little bit in blogs, but until now, no specific standard features are confirmed or solidly recommended due to the differentiation in the properties of text in each context. In this paper, we address the above issues by applying authorship identification on online diaries corpus using a different type of linguistic features and analyze those factors that affect the identification results.

The remainder of the paper is organized as follows. In the next section, we review the existing related work in authorship identification. The two following sections describe the nature of the language used in diaries or personal blogs and the utilized feature sets. Our main work follows in section 5, with the proposed framework and experiments. Results and discussions come next. Finally, the paper is concluded, and future work is also highlighted.

2 Related Works

Early work on authorship identification, on the Federalist Papers, is back to 1964 [11]. In this early work, a set of function words, which were not topic-related features, were utilized. Since then, Authorship identification has been researched in various text domains, such as emails, forums, and books as discussed below.

De Vel analyzed stylistics attributes to discover forensics in emails [5]. Although they achieved relatively good results, this may not be applicable straight-forward on the blogs due to the different nature of the text in emails and blogs. Generally, email text is shorter than diaries text and it is usually a topical dialogue between two authors, while online diaries text is from the author to the public, at least the intended group. Moreover in books and literature, Gamon [6] utilized the part-of-speech (POS) tri-grams and other features to find out the correspondent author out of just three writers. The main differences from our work are; the smaller number of authors and the nature of book text. Text in books is normally too long compared to text in diaries. And usually, there is a specific topic in the book. Books are also expected to be well written and proof read, which results in much less grammatical and syntactical errors than the case in personal blogs.

In the domain of web forums, Abbasi and Chen [1] used a collection of lexical, syntactical, structural, and content-specific features to find out the extreme patterns of writing on web forums. It may look that the text in web forums is similar to that in the personal blogs, but regularly there is a subject to be discussed in the forum, which in contrast to diaries that contains usually general ideas and thoughts on various and mixed issues.

Recently, the "Writeprints" technique was introduced in the domain of authorship identification [2], which separately model the features of each individual author by building the writeprint using the author's key features, instead of using one model for all the authors. Authorship attribution was also manipulated in probabilistic approaches using Markov chains of letters and words [16]. The above two methodologies are different in which they need to build an individual model for each author instead of just one model that classify all the authors. Although one model for each author will best represent the author style, this requires comparing the features from the new text against all the authors' models rather than testing through just one classification model.

The most common in all of above related works is that they have been developed for other types of text, other than personal blogs, which have their own properties as described in the next section. But to the best of our knowledge, authorship identification in personal blogs appears to have had less attention in literature. Gehrke et. al. [7] used Bayesian Classifier for each author, utilizing bi-grams word frequencies. In this work, all the posts from one author were combined in one document, as a bag-of-words model, for training and testing. In our work, we manipulate each post individually and build its features vector to be involved in training and testing process as described in details in framework design section. In addition to the difference in the utilized features, we build a single model for all the authors, instead of one model for each one.

From the above, it can be seen that author identification in personal blogs or diaries has received little attention. Consequently, no specific standard features are confirmed or solidly recommended due to the differentiation in the properties of text in each context. In the work presented in this paper, we address the above issues by applying authorship identification on an online diaries corpus using a different type of linguistic features and analyze those factors that affect the identification results.

3 Diaries Language

The style of writing in diaries blogs is different from other types of text such as emails, books, or articles. In this section, we briefly describe the nature and the properties of the language in online diaries. The text in online diaries is less focused and directed than other media. It contains thoughts, everyday stories and experiments, feelings, and opinions. The nature of personal diaries contains the personal print, details of blogger's life, and his or her experience. This type of text is rarely found on other corpora. The text in news columns might look similar to personal blogs as it comments about an event, opinion, or experiment, but usually in diaries, there is

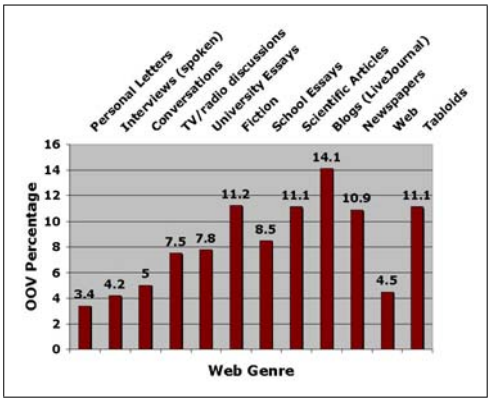

Figure 1: Out-of-Vocabulary percentages: personal blogs compared with various web genres

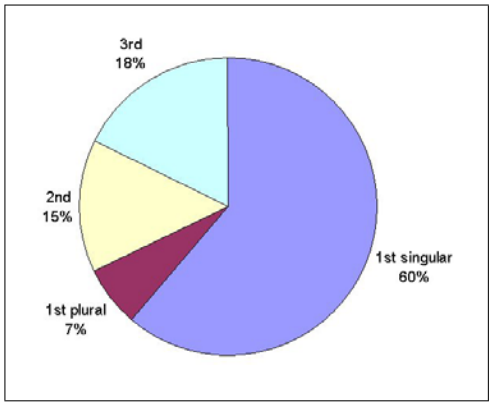

Figure 2: Pronouns usage in the corpus

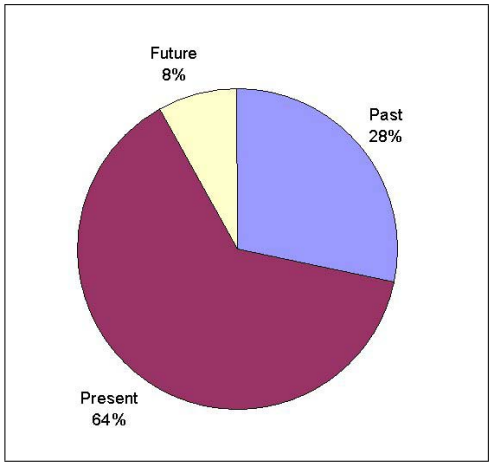

Figure 3: Tense usage in the corpus

no pre-determined subject or criteria for specific readers as in news text. Again as previously mentioned, diaries blog posts are different from emails as they are not written to a dedicated person, but it is available publicly to be accessed by everyone, sharing problems and ideas with friends and others. The authors are publishing their own diaries and they are more likely to use the words that express their feeling, mood, opinion, and emotions, at least from their point of view and according to their writing style. In writing diaries, people tend to use the everyday language and be less formal. Our selected text is challengeable as it is informal, self referential, combining spoken and written English, and rich of unedited content.

Mishne, in his study of the language of personal blogs [10], compares the personal blogs (Live Journal) with other types of web genres regarding the out-of-vocabulary (OOV) rate. OOV is measuring the percentage of new words that appear in testing and are not exist in training. Figure 1 shows a high OOV percentage in personal blogs which emphasize less focusing on a specific topic. This complexity of text motivates us to search for the best features that capture the style of user.

Furthermore, the language of personal blogs contains useful markers of personalities, emotions, cognitive, and social state [4]. People characteristics could be discovered from their language use. For example, young people will use more first person singular pronouns when they are under pressure, a greater sense of community when they include references to other people in their diaries, discard using present tense, include more articles and longer words when they are writing with high psychological distance [4].

Figure 2 and 3 show the usage percentage of Pronouns and Tense, respectively, in our selected corpus, extracted from the "LiveJournal" blog. In figure 2, we can see the high percentage of using the first singular pronoun (e.g., I, me, mine) in contrast to using the first plural (e.g., we, us, our), second singular/plural (e.g., you, your), or the third singular/plural pronouns (e.g., he, she, they, his, her). Figure 3 also indicates that the most dominant tense is the present tense, followed by the past tense, then the future tense.

These results agree with the type of writing in our corpus. As the authors are writing their own diaries, the use of the first singular pronouns is dominant. Also, as authors are usually writing diaries about their everyday activities or events, they are more likely to use the present tense. These characteristics require new types of features that can discriminate the style of the author. The following section will explain in details the selected features for this investigation.

4 Feature Set

A very important concern in text classification is the selection of features. In our investigation, we chose LIWC the Linguistic Inquiry Word Count [14], MRC Psycholinguistic database [17], and a collection of syntactic features. The majority of the features that have been selected have psychology basis, and known to be well related with the author's style and/or personality [9]. The

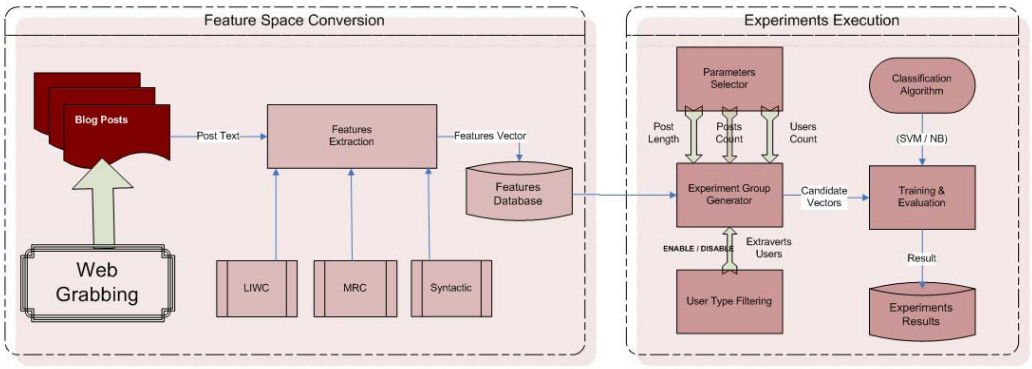

Figure 4: Authorship Identification Framework

properties of diaries text as they contain lots of feelings, personal activities, and thoughts are more captured using our selected features sets. The selected 63 LIWC features are grouped into four types:

1. **Standard linguistic features** (e.g., total word count, word per sentence, pronouns, punctuations, articles, time)
2. **Psychological features** (e.g., affect, cognition, biological processes)
3. **Personal concerns features** (e.g., work, sports, religion, sexuality)
4. **Paralinguistic features** (assents (e.g., agrees, ok), fillers (e.g., err, umm), non fluencies (e.g., I mean, you know))

In LIWC, the features are more of categories based on their intuitive meaning, including psychology and affect. These features (or categories) are evaluated by calculating the scores from a number of related words that are defined in the LIWC dictionary [14]. This means that the calculated word frequency is not used directly, but rather contributes to the final scores of multiple LIWC features. For example, the word "cried" is contributing to the calculation of the scores of five features: sadness, negative emotion, overall affect, verb, and past tense verb. Moreover, the LIWC can handle the different stems of the word, which is one of the common issues in natural language processing NLP. So the stem *hungr* captures the words *hungry, hungrier, hungriest* and so on.

The MRC database contains psycholinguistic statistics for more than 150,000 words. It includes frequencies among the lexicon such as: number of phonemes, number of syllables, imagebility rating, letters count, part-of-speech information, and familiarity rating. The syntactic features count the number of words and sentences, the frequencies of punctuations, abbreviations, and the usage of different types of the online text shortcuts.

It is worth mentioning that the LIWC has been used before in various contexts of linguistic analysis. It has been used on a text analysis task to obtain the personality values [15] according to the Big Five psychology model [12]. In that analysis, the LIWC features were extracted from students' essays, which is relatively more formal than blogs and manually tagged with personality values. Moving to personality recognition from text, LIWC features alone [8] and then together with MRC features [9], were utilized to investigate the personality factors values of the author. For text classification in particular, they have been utilized but for a limited number of classes, such as gender and/or age [13]. However, in authorship identification, the number of classes, users/authors in this case, is usually expected to be larger. In this investigation, we tried to study the use of the selected linguistic features with larger numbers of classes, representing users in the blog.

5 Main Work

In this section, we present the overall design and framework of our investigation as follows:

5.1 Framework Design

In this sub-section, we describe the design of the framework and the experiments for identifying the authors of blog posts. After grabbing the data corpus from the web, the extraction phase converts each post to a features vector containing the corresponding features values. This changes the input data from unstructured text space into features vectors space. All the vectors are stored in a database so that the manipulation of the features in the experiments is faster. The setup of our framework is depicted in figure 4.

First, we divided the input features vectors into groups according to three parameters: the post length, the number of authors, and the number of posts per author. Each group is manipulated individually by the classification

algorithm. In our framework we selected two machine learning algorithms: support vector machine (SVM) and Naive Bayes (NB). We depend mainly on SVM as the classification algorithm which is one of the best algorithms in this domain. We made a comparison between NB and SVM in the speed and accuracy as being described in following section.

For each experiment's data group, SVM is trained and tested by applying 10-fold cross validation. This means that there are 10 cycles of validation and the identification accuracy will be calculated among the average of them. In each cycle, 90% of the dataset are used for training and the remaining 10% are used for testing. We selected the implemented SVM algorithm (SMO) in the WEKA toolbox with linear kernel [18] for machine learning algorithms in our framework.

We choose 8 different numbers of authors, five different post counts per user, and 11 different post lengths. This makes 440 groups in total. Although there are 440 conditions to generate different vectors groups, for each condition, there are many candidate groups that satisfy it. For this reason, each experiment group is repeated 150 times, to handle as many combinations as possible of the different vector groups and calculate the overall average. Due to our limited corpus, few groups seemed to not have enough data satisfying some of the conditions. This reduced the total to 301 data groups, instead of 440. Hence, 45,150 experiments were executed using the support vector machine.

One of the main contributions of this paper is to study the effect of pre-filtering the candidate authors that are selected in the sampling stage of the classifier. In this study, we present the feasibility of building a classifier that contains the users which have common attributes such as personality properties. We try to find the type of personality either extraversion or introversion that is more correlated with authorship identification.

5.2 Corpus

We downloaded from LiveJournal 17,647 blog posts for 93 authors, with 200 posts as an average for each author. Although the text contains slang and shortcuts, no manual text pre-processing or filtering has been made over the posts, but an HTML stripping process was utilized to remove images and extract text from tables. This produced purely text documents to be used in our analysis.

6 Results

In this section, we present the results of our investigation. It should be mentioned that having three parameters investigated simultaneously, the result would ideally need to be represented in a four dimensional space. However this may not be easy to view/perceive. So, figure 5 depicts a selected 3D cube that represents the identification results according to the number of users and the post length. Similarly, figure 6 depicts a 3D cube representing the corresponding identification results according to the post length and the post size per user.

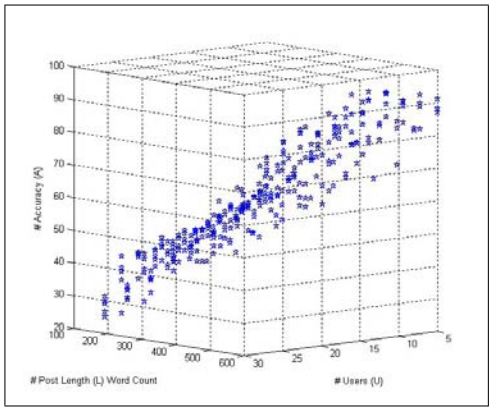

Figure 5: Identification accuracy(Users/Post Length)

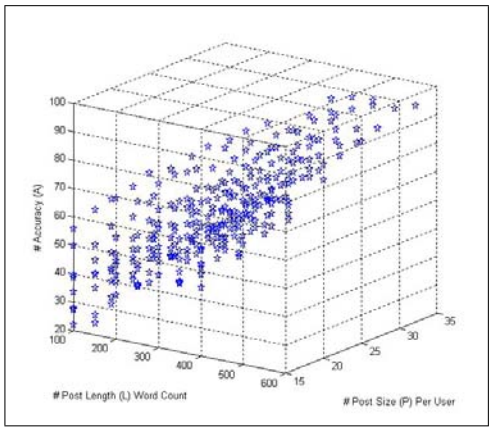

Figure 6: Identification accuracy(Post Size/Post Length)

The results, as presented in figure 5 and 6, justify the effective parameters ranges in which the identification percentage is more accurate. The two figures indicate that the identification accuracy is enhanced when there are more words in the post (post length). Although the selected features are less effective in short posts, having more posts (posts size) improves the identification accuracy as it provides more text written by the same author with different styles and contents, which is in turn included at the end in the learning process.

Generally, in SVM, there is a decline in the classification accuracy when the number of classes getting larger. We can notice that the identification results are higher when the number of users is between 5 and 11. Table 1 shows the difference between two ranges of user numbers among different post sizes and lengths. The threshold of user numbers has been selected according to the empirical boundaries we found in the number of users. We

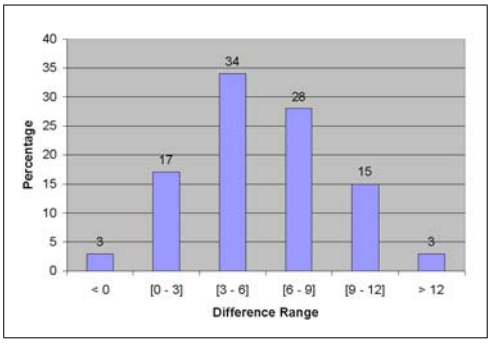

Figure 7: Percentage of accuracy difference between SVM and NB divided by ranges

achieved, as an average, 86% identification accuracy in specific ranges. It may also be noticed that some of the results do not exist because there was no enough available data for the corresponding ranges.

Table 1: Summary of SVM classification results with overall average of accuracy comparing two ranges of users' numbers

Length	Users <= 11	Users >11
	Average	Average
100	52.46	32.40
150	58.37	39.79
200	61.69	46.51
250	64.23	48.11
300	70.32	54.54
350	71.65	54.15
400	75.92	53.66
450	83.42	67.17
500	82.37	66.98
550	82.59	-
600	86.35	-

6.1 Comparison with Naive Bayes

In addition to the support vector machine, we applied the same experiments groups using the Naive Bayes (NB) classifier instead of SVM. We obtained relatively good results, but in most of the cases, SVM is outperforming NB, as expected. In figure 7, the difference ranges scoring between the two algorithms could be seen, among all the experiment groups. The difference average is 5.79 with 3.3 as the value of the standard deviation.

Table 2 shows the full result of testing authorship identification. Again, like SVM, we found that the identification percentage is highly different regarding the number of authors. So we can see two columns comparing the result when the number of users is less than or equal 11 and when the number is greater than 11. The results are calculated in the full range of posts size for each number of words (post length). Because we do not have a dataset for some parameters combinations, some of the cells in the table do not have value. A very important point to be mentioned here is that among all the experiments NB has been much faster of more than twenty times than SVM. The average experiment time for SVM is 12943 seconds while the NB takes only 619 seconds, in average. If the classification accuracy is the first priority, then SVM is the first choice. But when we have an autonomous system where the learning process is almost continuous and the speed is an important factor, and this range of difference is acceptable, NB could be the best choice or at least a compromise.

Table 2: Summary of NB classification results with overall average of accuracy comparing two ranges of users' numbers

Length	Users <= 11	Users >11
	Average	Average
100	46.41	30.09
150	50.25	35.23
200	58.83	42.41
250	61.14	45.78
300	61.65	47.56
350	64.72	49.45
400	68.17	50.97
450	73.47	57.74
500	76.61	61.55
550	76.24	-
600	77.32	-

6.2 Common Users Classifier

One of the big problems in authorship identification is to identify the author among large number of authors. Building different classifiers according to the type of users will decrease the number of the potential authors to be involved in each classifier. This would help in scaling the solutions with the increase in the number of authors. In this sub-experiment, we built a separate classifier for those authors who have similar personality attributes.

Writing diaries to be read publicly and describing the details of the private life to everybody on the internet is an indication that the bloggers are Extraverts [13]. Extraversion is one factor of the Big Five personality traits model [12]. The extravert person could be described for example as sociable, assertive, friendly, and playful. Another suggestion is that the bloggers are introvert because they are writing using nicknames on the blogging site, hiding their real identity [13].

We chose to test the authorship identification for those who are extraverts in their text. Although the corpus does not contain any tagging for extraversion, we extract

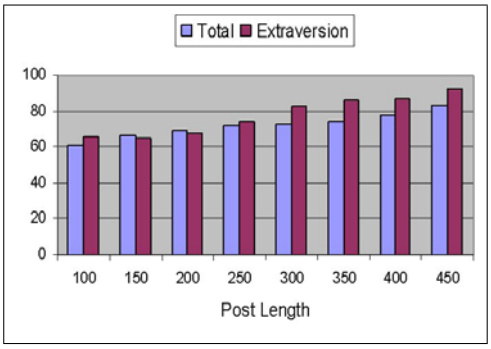

Figure 8: Classification accuracy for five users comparing the extraverts with the total users

the extraversion value automatically using a personality recognition software system [2] which computes estimates of personality scores along the Big Five dimensions. The lowest and the highest extraversion score is 1 and 7 respectively. We extract the corresponding value for each post and repeat the same previous experiments for the posts which have high extraversion values between 5 and 7.

SVM has been selected for these experiments. The extraversion condition filters the available posts and reduces the probable combinations according to the three parameters. We set the number of users to 5 with 3 different numbers of posts (15, 20, and 25) in 8 ranges of post length. Figure 8 displays the classification accuracy average for five users in the different post lengths between the extraverts and the total users. The results indicate that those who have a high extraversion score are better classified in the authorship identification process. This motivates us in future to find more user properties which can be utilized to have a multi-classifier hierarchy that includes several classifiers for several users' types.

7 Conclusion and Future Work

In this paper, we presented our investigation of identifying the bloggers in online diaries by mining the diaries text of each blogger. The investigation contains majorly three contributions. The first one was by utilizing two psycholinguistic features, namely the LIWC and MRC feature sets together, for the first time on the personal blogs for blogger/authorship identification. The second one was the analyzing of the effect of various parameters, and feature sets, on the identification performance. This included the number of authors in the data corpus, the post size or the word count, and the number of posts for each blogger. Finally, we studied the identification outcome for shared-attribute authors. While previous studies in authorship identification achieved high classification accuracy but in different corpus types, we also acquire, according to specific criteria, superior results using a smaller number of features (102 features), compared to their features numbers. The design of the authorship identification experiments framework allows evaluating different types of machine learning algorithms using several forms of features and comparing the results over large numbers of the experiments combinations.

The selected feature sets have been confirmed to identify the author style in personal text, over multiple documents. We studied the effect of each of the selected three parameters, as well as the filtering stage, on the identification accuracy. We found that the post length, or the number of words in the text, is highly contributing to the author style attribution. Having more words facilitates more accurate and stable identification performance as the author style can be more appropriately captured. The results provided the preferred ranges of those parameters, which can be used as recommendation for further studies in authorship identification in personal blogs.

In addition, although we achieved relatively good results with the selected features, we are planning to test the system with a subset of the same features searching for the best feature set that can better discriminate the authors. Our initial results in testing the common users classifier with the filtering stage of extravert authors are promising to search for other criteria in future that can decrease the large number of authors, and better produce different classifiers for several users' properties. This experiment runs over automatically generated values for extraversion. We thought to test the system using other types of corpus which contains tagging like personality traits or in advance the gender. This aims to study the difference in accuracy between male and female authors. Moreover, the bloggers which are effective in the blog are different in their style from the inactive ones. Selecting those influential bloggers is not clearly related to the posting rank (i.e. not the number of posts). Instead other methodologies are now developing to specify this kind of influential people [3].

There are many tracks to be developed in this area of research. For example, the selected features do not cover all the properties of the blog like misspelling, shortcuts, and emphasizing words. We are currently developing new features to deal with those specific properties and we are interested in comparing the resulting accuracy and its improvement.

[2] http://mi.eng.cam.ac.uk/ farm2/personality/recognizer.html

References

[1] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE INTELLIGENT SYSTEMS*, pages 67–75, 2005.

[2] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similar-

ity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):1–29, 2008.

[3] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pages 207–218. ACM New York, NY, USA, 2008.

[4] M. A. Cohn, M. R. Mehl, and J. W. Pennebaker. Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science*, 15(10):687–693, 2004.

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

[6] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics Morristown, NJ, USA, 2004.

[7] G. T. Gehrke, S. Reader, and K. M. Squire. Authorship discovery in blogs using bayesian classification with corrective scaling, 2008.

[8] A. Gill. Personality and language: The projection and perception of personality in computer-mediated communication, 2003.

[9] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[10] G. A. Mishne. *Applied Text Analytics for Blogs*. Universiteit van Amsterdam, 2007.

[11] F. Mosteller and D. L. Wallace. *Inference and disputed authorship: The Federalist*. Reading, Mass.: Addison-Wesley, 1964.

[12] W. T. NORMAN. Toward an adequate taxonomy of personality attributes: replicated factors structure in peer nomination personality ratings. *Journal of abnormal and social psychology*, 66:574–583, 1963.

[13] S. Nowson and J. Oberlander. The identity of bloggers: Openness and gender in personal weblogs. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*, 2006.

[14] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway : Lawrence Erlbaum Associates*, 2001.

[15] J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–1312, 1999.

[16] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking. In *Proceeding of 2006 Confereance on Empirical Methods in Natural Language Processing (EMNLP)*, pages 482–491. Association for Computational Linguistics, 2006.

[17] M. Wilson. Mrc psycholinguistic database: Machine usable dictionary. *Information Division Science and Engineering Research Council*, 1987.

[18] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2005.