2009 International Conference on Computer Engineering and Applications
IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

More Blogging Features for Author Identification

Haytham Mohtasseb[+] and Amr Ahmed

Department of Computing and Informatics
University of Lincoln

Abstract. In this paper we present a novel improvement in the field of authorship identification in personal blogs. The improvement in authorship identification, in our work, is by utilizing a hybrid collection of linguistic features that best capture the style of users in diaries blogs. The features sets contain LIWC with its psychology background, a collection of syntactic features & part-of-speech (POS), and the misspelling errors features.
Furthermore, we analyze the contribution of each feature set on the final result and compare the outcome of using different combination from the selected feature sets. Our new categorization of misspelling words which are mapped into numerical features, are noticeably enhancing the classification results. The paper also confirms the best ranges of several parameters that affect the final result of authorship identification such as the author numbers, words number in each post, and the number of documents/posts for each author/user. The results and evaluation show that the utilized features are compact, while their performance is highly comparable with other much larger feature sets.

Keywords: Blogs Mining, Authorship Identification, Machine Learning.

1. Introduction

Blogs are one of the most popular forms of users' contribution to the web contents. There are many categorizations of blogs which are differing in the content, publishing methodology, and even in the type of readers[1]. Personal blog, or online diary, is the most famous category in which the blogger expresses his 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"[2]. 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 which is defined as the process of discovering or distinguishing the author of a given particular text from a set of candidate authors.

Author 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. Authorship identification has been applied 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. Moreover, there are many factors that have important roles and affect the performance of identification process such as the text length, the number of authors, the number of posts per author, and the type of authors. In this paper, we address the above issues by applying authorship identification on an online diaries corpus using a different type of linguistic features with numerous combinations and analyze those factors that affect the identification results. The remainder of

[+] Corresponding author Tel.: + (44 1522 886 162); fax: (44 1522 886 974). *E-mail address*: hmohtasseb@lincoln.ac.uk
[1] http://en.wikipedia.org/wiki/Blog
[2] http://www.livejournal.com

the paper is organized as follows. In Section 2, we review the existing related work in authorship identification. Section 3 describes our study of the text properties and the nature of the language in diaries blogs. The framework follows in section 4, with our proposed feature sets, experiments, and corpus. Results and discussions are in Section 5. Finally, the paper is concluded in section 6.

2. Related Works

Starting from works in email authorship identification, De Vel analyzed stylistics attributes to discover forensics in emails [3]. 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 short in length and it is usually a topical dialogue between two authors, while online diaries text is longer and it is from the author to the public, at least the intended group. Also in books and literature, Gamon [4] 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 et. al. [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, Abbasi et. al. [2] presented the "Writeprints" technique, which separately model the features of each individual author, instead of using one model for all the authors. They build a writeprint for each author using the author's key features. Authorship attribution was also manipulated in probabilistic approaches using Markov chains of letters and words [11]. 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. Koppel and Schler depend mainly on misspelling features in addition to other lexical and syntactic sets to identify the author in email text [7]. Although some of our features are similar to theirs, we have smaller number of misspelling error features (11 features) compared to theirs (99 features). With this compact number of features we achieved higher results in the corresponding ranges. Furthermore, the created misspelling features are highly correlated with the diaries text. We also analyze different ranges of user numbers and words count, addressing the effective ranges of those features.

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. [5] 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 sections 4. 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. Text Properties

The main target of this section is to illuminate our methodology of feature selection, according to what we found in this study of the text properties in blog diaries. 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 no pre-determined subject or criteria for specific readers as in news text. Again as previously mentioned, diaries 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 [8], in his study of the language of personal blogs, compares the personal blogs (LiveJournal) 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. He found a high OOV percentage in personal blogs which emphasize less focusing on a specific topic.

The complexity of text and the high percentage of new words motivate us to focus more on these new or misspelling words that could appear in the text. We found that a large percent of the misspelling errors came mainly from either emphasizing or naming words. As the text is not showing usually the current feelings and emotions of the writer in online world, users tend to create new types of text highlighting more what they mean. Emphasized words are commonly used in the internet by repeating one of the characters (coooool), capitalization (STOP IT), or by utilizing the editing tools such as making the text in bold or different colours. Figure 1 illustrates the extracted categorizations of misspelling errors in our corpus. It is clear that in addition to unclassified error type, a large percent of misspelling words is for emphasized words. Next, in the subsection of feature sets, we explain more how we get benefits from these text properties to find the most suitable textual features.

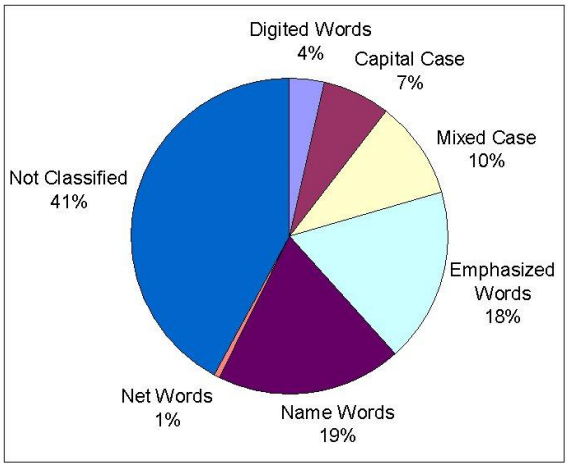

Fig. 1: Misspelling errors categorizations

4. Framework

In this section, we present our authorship identification framework illustrating the details of the utilized feature sets, text collection and corpus building, and the experimental work together with the framework design.

4.1. Feature Sets

In text classification tasks, in addition to the classification algorithm, features selection is more serious and plays an important role in the final results. In our study of the nature of the text in blogs, we utilized the best features that suit the author's style in diaries. We have totally 129 features which is a small number compared to other studies in the same domain. This section will explain more the utilized features according to their category.

4.1.1. *LIWC* : We chose LIWC the Linguistic Inquiry Word Count [9] as it has psychology basis, and known relate well with the author's style and/or personality [6][10]. The properties of diaries text as they contains 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:

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

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.

4.1.2. *POS & Syntactic*: The Features extracted from Part-of-speech (POS) tags are commonly used in text classification tasks. They describe more the syntactic structure used by the writer. As our corpus do not contain this tagging, we used Stanford POS tagger [12] to tag all the posts in the corpus. Then, we built up for each tag type, the corresponding counting feature. The syntactic features count the number of words and sentences, the frequencies of punctuations, abbreviations.

4.1.3. *Misspelling Errors:* Blogging text contains lots of spoken language words, shortcuts, and other words imported from different language rather than English. This may reflect the background, home country, and the previous experience of the blogger. We extract the misspelling error words from each post using ASPELL algorithm[3], classify the errors, using a set of regular expressions, into seven categories as depicted in Figure 1, and find the correction suggestions of each word. We used three versions of the ASPELL English dictionaries: the General English, the British, and the American to catch as most as possible English words. For finding the corrections, we rely on Levenshtein[4] string edit distance algorithm to find the suggested correction for the misspelled word. The distance between two words is the minimum number of operations (inserting, deleting, or replacing a character) needed to change one word into the other word. The extracted features are finally representing the counting of the different errors categories and classifying the number of suggestions into different ranges. From these ranges, we particularly give more weight for the count of error words which have just one suggestion as it will be corrected directly.

4.2. Text Collection

We downloaded from LiveJournal 80000 blog posts. This includes 565 authors with 140 posts as an average for each user. The total number of words is 20,172,275. After HTML stripping process which removes images, videos, extract text from tables, and delete empty posts, we finally have a corpus that contains 63167 posts. This produced a corpus of purely text documents to be used in our analysis.

4.3. Experimental Works

In this section we explain in details the stages of authorship identification framework that starts with text collection, as described in the previous subsection. Next, in features extraction stage, every blog post is converted to a features vector, storing the values in a relational database which simplify and increase the speed of all the ongoing experiments. Because we need to test the system for different ranges of parameters (number of users, number of posts per user, and number of words per post), we decided to divide the vectors into groups according to those parameters. We chose six different numbers of authors, five different post counts per user, and eleven different post lengths (words number per post). This makes 330 groups in total.

[3] http://aspell.net

[4] http://en.wikipedia.org/wiki/levenshtein_distance

Although there are 330 conditions to generate dissimilar 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 finally calculate the overall average. We select SVM as the classification algorithm which is one of the best algorithms in this domain. 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 cycles. In each cycle, 90% of the dataset is used for training and the remaining 10% is used for testing. We selected the implemented SVM algorithm in the WEKA toolbox with linear kernel [13]. In the design of the framework we allow the use of separate features set in the identification process. The candidate features vector could contain the values of any combination from the original feature sets. In section 5.1, we present the results of repeating the same work above for numerous combinations from the features set and compare their results among different parameters.

5. Results and Discussions

Using our proposed framework, we found that the identification results vary according to the parameters ranges. The post length or the number of words in each post is playing an effective role in the final outcome. As shown in figure 2, we achieved a high classification percentage for larger size posts, exceeding 90% for some ranges. Although having more words will enhance the result, a minimum of 250 words as an average post length could be essential to capture the identity of the author. One of the key problems in SVM is the large number of classes (users in this case). Figure 2 signifies the variations by showing different trends according to the number of users. As the number of users grows, the result drops down. In the other hand, while the total number of features is small compared with other studies, we achieved high identification results when the number of users is less than twelve. However, for the remaining ranges, the results are still commonly around 70% as an average result. Finally, most of the experiments outcomes are greater than the baseline (50%). One more thing, which is not presented explicitly, is the role of the number of posts per user. We found that having more posts is enhancing the final results. As more documents are being involved, the model will better represent the user by including the different styles, emotions, and contents of the blogger. In the next subsection, we present more analysis of the utilized features and their contributions to the final result.

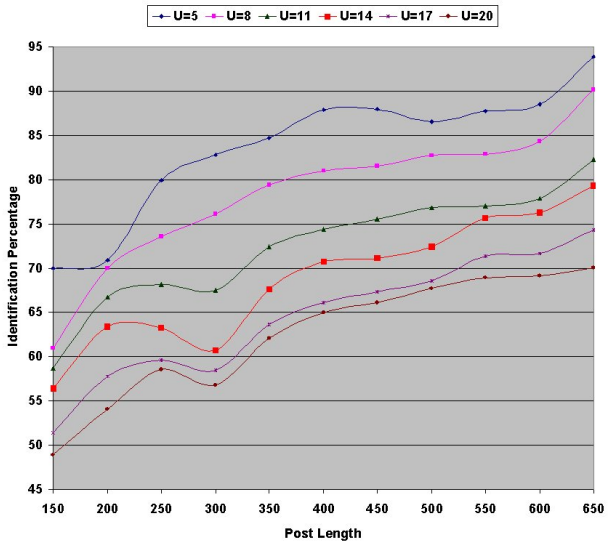

Fig. 2: Identification percentage for different users count according to the post length

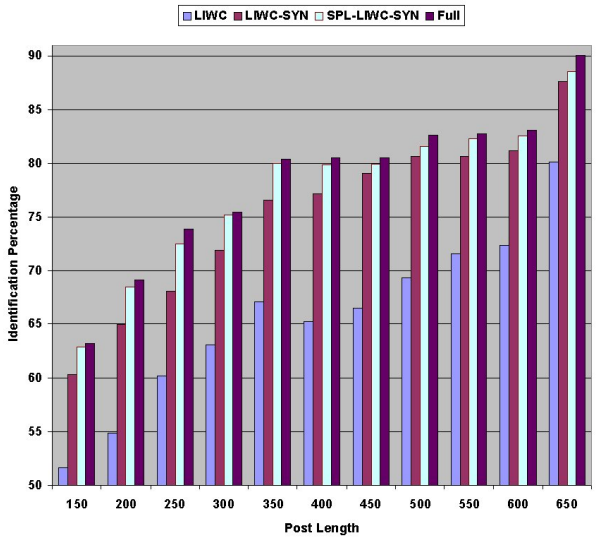

Fig. 3: Identification percentage for several feature set combinations according to the post length

5.1. Features Comparisons

In order to analyze the selected feature sets in our problem, the same experiments were executed with different combinations of feature sets. In these experiments we found that LIWC alone was the best among all the other feature sets, when each set is used individually. It has provided a good classification result in all ranges, up to 80% for 650 words. This is due to its rich dictionary which is covering different topics and backgrounds. We chose it as a baseline and accumulatively add other feature sets, and compare their results, as displayed in figure 3. Adding the syntactic features set to LIWC is significantly improving the results comparing with other options. Misspelling errors features have also an effective role in enhancing the

percentage especially in lower post lengths. One of the main contributions in this paper is having this enhancement in the results with our misspelling errors categorizations features which are highly associated with blogging text attributes. In contrast, adding POS, forming the full features, has a little effect compared to the other feature sets.

6. Conclusion

In this paper, we presented our research of identifying the bloggers in online diaries by mining their diaries text. We identify the nature and properties of the textual content used by bloggers and find out the superlative collections of linguistic features that best capture the style of authors. In our framework, a large spectrum of experiments have been executed, exploring the significant parameters ranges of the users' number, posts sizes and lengths, and indicating the best set of features that improve the identification percentage. 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 (129) compared to their features numbers.

We found that LIWC is the best individual option among other feature sets as a baseline selection. This is due to its dictionary richness which covers a large variety of real life topics that is highly correlated with the content of the diaries blogging text. In additions to the other features sets, the syntactic & POS, which are also improving the result, our created set of misspelling features is enhancing the final outcome of the authorship identification framework. Although previous studies utilized misspelling features, but we chose a very small number of features than their features size, considered the common misspelling errors happened in the diaries, and effectively introduced a new categorization map between the features and the misspelling words.

7. 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 similarity detection in cyberspace. ACM Transaction Information Systems, 26(2):1–29, 2008.

[3] 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.

[4] M. Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Proceedings of the 20th international conference on Computational Linguistics, 2004.

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

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

[7] M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, pages 69–72, 2003.

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

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

[10] 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, Dec 1999.

[11] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, Markov chains and author unmasking. In Proceeding of 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP), page 482491.

[12] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter on Human Language Technology, volume 1, pages 173–180, NJ, USA, 2003. Association for Computational Linguistics Morristown.

[13] I. H. Witten and E. Frank. Data mining: Practical machine learning tools and techniques. 2005.