Content-centric age and gender profiling
Notebook for PAN at CLEF 2013

Wee-Yong Lim, Jonathan Goh and Vrizlynn L. L. Thing

Cybercrime and Security Intelligence (CSI) Department
Institute for Infocomm Research, Singapore
{weylim, jonathan-goh, vriz}@i2r.a-star.edu.sg

Abstract Author profiling can be considered a form of text analysis of which the objective is to ascertain characteristics of the author behind a text sample. This paper describe the design and implementation of an approach for determining the age group (10s, 20s, or 30s) and gender (male/female) of text samples for the author profiling task in PAN 2013. Evaluation is then based on the compounded accuracy in determining the correct age group and gender of authors of samples in a test corpus. The training corpus provided for this task contains English and Spanish text samples from online contents (e.g. blogs, chats) of authors. Content in each sample are split into one or more "conversations", of which are all wholly attributed to a specific author. To the best of our knowledge, interweaving responses of other person(s)(if any) are filtered, focussing the scope of the analysis to the writing style and content present in individual author's sample. The underlying research in this work is, thus, the empirical investigation of features that can be extracted from the text samples, that are helpful in identifying the gender and age group of an author based purely on characteristics present within his/her text samples.
Main contribution in this work is a concise content-based feature based on similarity scores between given text samples and corpora of the different classes. This feature is compared and used with some common style-based, vocabulary and idiosyncrasies features. Results from experiments on a balanced subset of the PAN 2013 authorship profiling training corpus paint a clear contrast between the content-based feature and the other features, favouring the former for both the English and Spanish samples. Ultimately, 24 five-fold cross validation tests were ran on the different feature sets on the balanced corpus, with the best accuracies for simultaneous gender and age group classification at 48.52% and 61.23% for the English and Spanish samples respectively, in contrast to a baseline of 16.67%.

1 Task and General Approach

Text analysis can involved processing users' generated content for various purposes such as classification/clustering-based tasks like author attribution [7][10][4], plagiarism detection [12] and information retrieval related tasks such as information extraction, summarization of contents, etc. This work focus on the task of profiling the background/characteristics of groups of authors by analysing their text samples. In particular, the premise of this work is concerned with the author profiling task in PAN 2013 [1]

- profiling the age group (10s, 20s, or 30s) and gender (male/female) of authors using a provided training corpus.

Similar to the task of author attribution, it is recognized here that the key research in author profiling is the selection for the best features and the proper use of classification techniques in building an appropriate model to distinguish between the different profile groups. Thus, in approaching this problem, we seek the question *"Why/How do authors in different sociolinguistic profile group differs in their written communications, assuming using a common language?"*. In general, there are 2 main contributing factors to differences in the communications amongst the authors in the different groups - (i) content/subject matter difference as well as (ii) syntactic and style-based differences [2] amongst the different profile groups.

Profiling text samples typically contain the sequential steps of describing the text sample (usually via features represented in a vector), investigation and selection of useful features and lastly, building a model to represent characteristics styles of each author or author group. In this work, we apply Principal Component Analysis (PCA) to linearly transform the high dimensional data into a lower dimensional space for a more simple representation of the data and subsequently utilize a popular implementation of Support Vector Machine (SVM) [3] classifier for learning the model for the author profiling task.

The rest of this paper is organised as follows. Section 2 describe the motivation behind our approach to the PAN 2013 author profiling task, ending with a tabular listing of the features used. Section 3 present the results of the experiments in this work together with analysis of the various features used. Section 4 summarize the approach used and findings of our analysis in this work.

2 Our Features

A person's syntactic construct or lexical usage can give cues to his authorship, but what features do we use to describe and quantify such characteristics? Despite several prior research on authorship attribution [9], [10], [11], [7], [13] and some on author profiling [2], [5] [8] , there is little consensus on the "ideal" features to use [7]. Hence, our approach is to combine the use of common or relevant features used in previous work with a content similarity feature which we have implemented for this task.

2.1 Style-based features

Analysis on empirical evidence indicated the usefulness of pronouns, determiners and prepositions (e.g. "I", "her", "as", "the", "of", "in", etc.) in gender and age group profiling tasks in [2]. Hence, features of similar intent in the form of POS tags and specific pronouns are used in this work to (i) corroborate the usefulness of such features for the shorter mean sample length corpus and (ii) explore their usefulness on a Spanish corpus as well.

Other simple statistical features such as average sentence length, words per conversation, number of contraction words (e.g. he's, i'm, etc.) and number of URLs are also included as part of the style-based features. Given the relatively short length of the

text samples, all features are simply normalized against the number of words in each sample.

2.2 Vocabulary and Idiosyncrasies (VI) features

The number of unique words in the sample is used as a proxy for the vocabulary richness of the author of the sample. This is a simple and intuitive feature but further consideration may be needed in future work in view of the discussions and caution raised on the use of vocabulary features in [6], [7].

Referencing the list of discriminative frequent words in [2], we postulate that in some cases (e.g. the teens age group), the presence and occurrence of neologisms can help in discriminating between the different profile classes. Hence, apart from the nature of a set of particular words themselves that is discriminative, idiosyncrasies represented by such words may that help in this task, where the term "idiosyncrasies" in this work refers loosely to any (misspelled) words that deviate from a standard US/GB English or Spanish dictionary lexicon or any all-caps words with 2 or more characters.

2.3 Content-based features

In this work, we consider content-based features as features that are reflective of the subject areas expressed in the samples. A straightforward measure for generating such features for a certain author group profile is via a histogram of the n most common words found in training samples from the group. Naturally, this concept can be extended to character or word level n-grams in the samples. Two inherent practical weaknesses are present in such a method - (i) high number of dimensions in the description and (ii) a thresholding exercise is required to determine the best n to use. In addition, a thresholding exercise also implies that there is likely to be some form of information loss. Intuitively, this is represented by the loss of not-so-frequent, yet still discriminative helpful terms that happen to fall below an arbitrarily chosen threshold.

For practical reasons, the approach taken in this work seeks to minimize the dimension of the description for each text sample and eliminate the need for any thresholding exercises. As such, term frequency-inverse document frequency (TF-IDF) based scores are used to measure the similarity of content between a given text sample and each of the profile group's collection of samples. This is done by measuring each word of the text sample with the entire lexicon for each of the profile group. The TF-IDF scores of all the words are then summed up to determine the similarity between the text sample and the corpora. Ideally, the collection of training samples from the correct profile group will give the highest similarity score, indicating a similarity in the content present in the given text sample and training samples from its profile group. In this work, we chose to use the similarity measures from all the profile groups as a feature set.

In order to obtain the similarity score, the TF-IDF for all the words in in the entire training corpus are first calculated, thereby providing an indication on how rare or common a particular word is in the entire corpus. TF is first calculated by normalising the term frequency (TF) over the maximum frequency in the corpus as stated in *Equation* 1. This is to prevent the TF from being bias towards the profile groups with larger lexicon set.

$$tf(t,c) = \frac{f(t,c)}{max(w,c)} \tag{1}$$

where *f(t, c)* refers to the frequency of the word *t* in the corpus, *c* and *max(w, c)* refers to the maximum frequency of any word, *w*, in the corpus, *c*. Subsequently, the TF-IDF score for each word is then computed using *Equation* 2.

$$TF\_IDF_c = tf(t,c) \times log(\frac{sum\_of\_freq}{tf(t,c)}) \tag{2}$$

where *tf(t, c)* is simply the normalised frequency of the word that occurs in a particular corpus, *c*. *sum_of_freq* is defined as the sum of frequency for the entire corpora. The similarity score from a test sample would then be calculated using *Equation* 3.

$$SimilarityMeasure_c = \sum_{i=1}^{n} (1 + log(\frac{term\_freq(w)}{n}) \times) TF\_IDF_c(w)) \tag{3}$$

where *SimilarityMeasure*$_c$ refers to the score between the test sample and the corpus, *c*. *n* refers to the total word count in the test sample, *term_freq* is defined as the frequency of the word, *w*, in the test sample. *TF_IDF*$_c$ is the pre-determined score for the word, *w*, in corpus, *c*.

2.4 Features list

Features in Table 1 summarize the features used in this work. Values are normalized against length of each sample.

3 Experiments

The English corpus for the PAN 2013 author age and gender profiling consist of 236,600 samples equally divided between the genders. Number of samples for the 10s, 20s and 30s age groups are 17,200, 85,800 and 133,600 respectively. The smaller Spanish corpus consist of 75,900 samples also equally divided between the genders and having 2,500, 42,600 and 30,800 samples for the respective age groups. For each language, classification can be performed separately or simultaneously among the two sets of classes, resulting in a total of 6 classes, namely [*male, 10s*], [*female, 10s*], [*male, 20s*]... etc.

In our experiments, we seek to quantify the discriminative capabilities of the different sets of features for the *age*, *gender* and *age+gender* classes. This results in a total of 12 experiments being conducted for each language. Each experiment consist of a coarse, but reasonably extensive search for the best parameters for the SVM classifier and the corresponding 5-fold cross validation accuracy. To avoid the problem of having an unbalance dataset, 1,000 samples were selected for each of the *age+gender* classes and subsequently 2,000 samples for each of the *age* classes and 3,000 samples for each of the *gender* classes.

Feature	Description
Style-based	
Average words	Average number of words per conversation (each sample contains one or more 'conversation' by the same author)
Contraction words	Number of abbreviated words using single quote marks
Average sentence length	Average number of words in each sentence
URLs	Number of URLs
Punctuation list	Histogram of punctuations
Pronoun list	Histogram of pronouns (e.g. *they, you, she, etc.*)
POS tag list	Histogram of Part-of-Speech tags (e.g. *adjective, determiner, etc.*)
Vocabulary & Idiosyncrasies (VI)	
Unique words	Number of unique words in the sample
Capital words	Number of words (length > 2) containing only capital letters
Spelling errors	Number of words that are not found in US/GB English or Spanish dictionaries
Content-based	
Lexicon similarity list	Similarity scores of the sample with the content of the different classes

Table 1. List of features

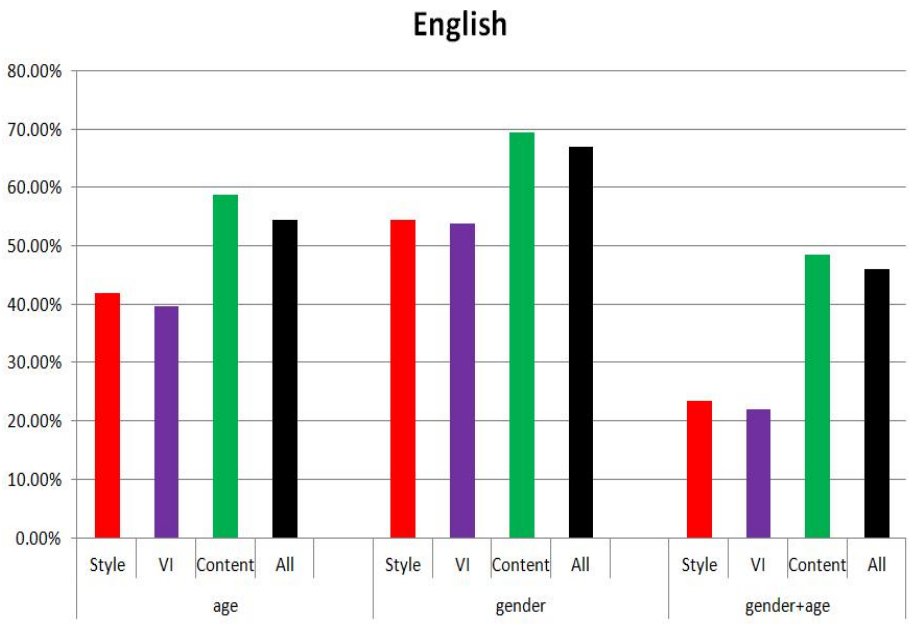

Figure 1. 5-fold cross validation accuracies for English

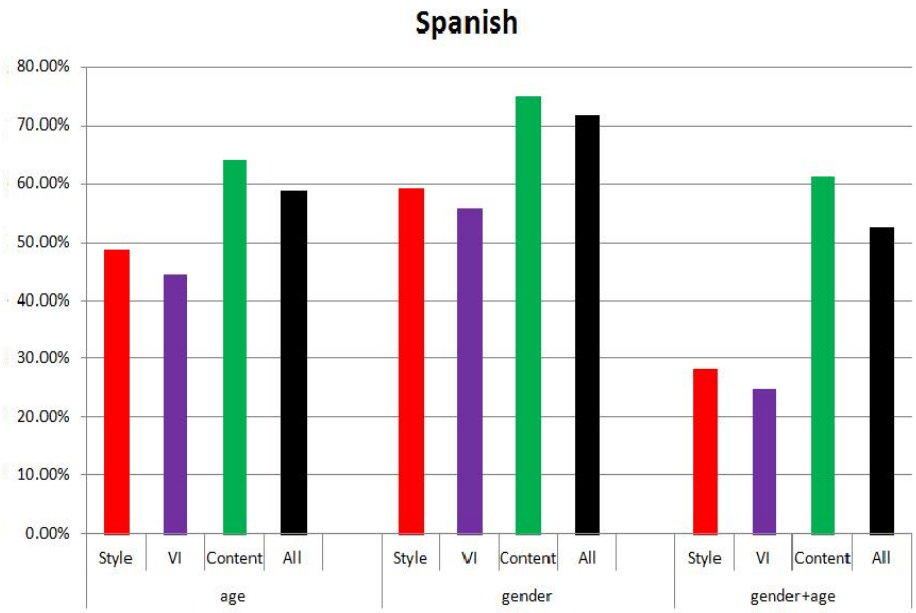

Figure 2. 5-fold cross validation accuracies for Spanish

Given the balanced dataset used in the experiments, the baseline for the *age*, *gender* and *age+gender* classifications are approximately 33%, 50% and 17% respectively. Referring to Figures 1, 2, it is clear that style-based, vocabulary and idiosyncrasy features are simply not very discriminative in discerning between the age and gender of the samples in this specific task. Their poor performance is apparent for both languages and is in stark contrast to the accuracies obtained for the content-based features.

In contrast, content-based features are much more discriminative than the other features for both age and gender profiling for the given corpus. In fact, the clear difference in performance is further highlighted by the slight *decrease* in accuracy when content-based features are used together with the rest of the features. Given the consistent results for different feature sets across the classes and languages, we are pessimistic on the use of the listed style-based, vocabulary and idiosyncrasies features in this work for age and gender profiling tasks.

The weak performance of the non-content features in this work may indicate that writing styles (i) may not be consistent enough across a community of author even if these authors share some sociolinguistic similarities in one form or another (ii) or that differences in writing styles are simply not evident enough in the relatively short sample length, (iii) or that the listed features are ineffective at capturing the relevant stylistic properties. Considering the stronger performance of rich set of syntactic features in other work (e.g. [8]), we are reluctant to dismiss the effectiveness of such style-based features. Furthermore, it is acknowledged that the motivation for minimizing the num-

ber of dimensions in this work inhibited the investigation of a more comprehensive set of style-based features.

The task of profiling authors using their writing samples is effectively categorizing the authors into their most likely sociolinguistic groups based on similarities of their samples with collective demographic characteristics exhibited by each these groups. It is then intuitive and apparent from previous work that a major component of such demographic characteristics is defined by popular subject areas in the respective groups. Given "many different topics" are present in the PAN corpus, it is assumed that there are no constraint in the subject areas in all the text samples, allowing authors to freely steer their conversations to the topics of their interest. It is, thus, of no surprise that content-based features subsequently performed relatively well in discriminating samples into their authors' age group and gender. Naturally, the imposition of any constraints based on subject areas when selecting the training and test text samples is likely to quickly deteriorate the effectiveness of any content-based features.

Caution have been expressed in the past noting the dependency of content-based features on the situation or experimental setup [2]. That is the performance of classifier(s) trained using content-based features is likely to be significantly negatively affected when used to classify samples whose subject area is different from those present in the training samples. However, it is argued that such caution is applicable to style-based (or other) features as well - the difference in both content and style can be attested to the different sociolinguistic characteristics of the profile groups. Hence, it is beneficial for researchers and investigators to exploit such characteristics when performing author profiling based on text samples.

4 Conclusion

This paper describe our approach for the author age group and gender profiling task in PAN 2013. Investigations on the discriminative capabilities of a set of style-based, vocabulary, idiosyncrasies and content-based features are conducted in this work. Although the non-content-based features performed poorly in our investigations, we are reluctant to dismiss their suitability for all author profiling tasks based on the limited range of the types of features implemented and the possibilities of these features being negatively affected by the relatively short sample length. On the other hand, the novel and concise content-based feature used in this task prove its effectiveness in discriminating the samples to their authors' gender and age groups. The approach to generate this content-based feature is applicable in all domains and its low dimensionality output will facilitate any downstream machine learning process.

References

1. Pan 2013 (April 2013), http://pan.webis.de/
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2) (2009)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

4. van Halteren, H.: Linguistic profiling for author recognition and verification. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (2004)
5. Halteren, H.V.: Author verification by linguistic profiling: An exploration of the parameter space. ACM Trans. Speech Lang. Process. 4(1) (Feb 2007)
6. Hoover, D.L.: Another perspective on vocabulary richness. Computers and the Humanities 37(2) (2003)
7. Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1(3) (2006)
8. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17, 401–412 (2003)
9. Koppel, M., Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: In IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis (2003)
10. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. 60(1) (Jan 2009)
11. Labbé, D.: Experiments on authorship attribution by intertextual distance in English. Journal of Quantitative Linguistics 14 (2007)
12. Oberreuter, G., L'Huillier, G., Ríos, S.A., Velásquez, J.D.: Approaches for intrinsic and external plagiarism detection - notebook for pan at clef 2011. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
13. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3) (2006)