Semantic-based Features for Author Profiling Identification: First insights

Notebook for PAN at CLEF 2013

Delia-Irazú Hernández[1], Rafael Guzmán-Cabrera[2], Antonio Reyes[3], and Martha-Alicia Rocha[1,4]

[1] Universidad Politécnica de Valencia, España
dirazuherfa@hotmail.com, mrocha@dsic.upv.es
[2] Universidad de Guanajuato, México
guzmanc@ugto.mx
[3] Instituto Superior de Intérpretes y Traductores, México
antonioreyes@isit.edu.mx
[4] Instituto Tecnológico de León, México

Abstract In this article we present a semantic-based approach concerning the identification of particular author's traits, such as age and gender, from social media texts. The model here described is intended to provide information on different levels of analysis: from textual markers to semantics. Different classifiers were used to assess the performance and scope of the model.

1 Introduction

Nowadays, social interaction through Internet is becoming a major problem due to the insufficient control regarding the authenticity of users profiles. For instance, a 35 years old man may easily impersonate anybody just by creating a fake profile. The consequences are in some cases very dangerous. In particular, if we consider pedophilia, bullying, extortion, etc.

Author profiling is the task of identifying personal characteristics of Internet users (such as age, gender, native language) based on analysing their interactions, mainly, considering textual patterns in their texts. The task has various applications such as security, forensics, marketing, among others.

In this article we present an approach for identifying two main characteristics regarding the way in which Internet users interact: age and gender. The approach is grounded on detecting textual features considering different types of information: from to textual markers to semantics. The article is organized as follows: Section 2 introduces the features of our model. Section 3 describes the set of experiments carried out to assess the model, as well as the main results. Finally, in Section 4, the main conclusions are given.

2 Features Description

This section describes the set of features used in our model. Each feature is intended to provide information concerning different levels of analysis: from textual markers to semantics. They are:

1. $Signatures$: concerning explicit linguistic markers within a text.
2. $Chatslang$: concerning words and expressions commonly used in internet forums.
3. $Context$: concerning the presence of discriminating clusters across the classes.
4. $Emotionality$: concerning the use of words to communicate emotions, feelings, moods, etc.
5. $Semantic\ similarity$: concerning the semantic relatedness among the words of a text.

As described in [4], signatures are intended to identify textual markers that are used to throw focus onto certain aspects of a text. For instance, the use of capitals or emoticons. Such elements are often used to communicate something implicitly. Let us consider the presence of words in capitals. Beyond their morphosyntantic category (noun, adjective, verb), such words may reveal underlying information that is not explicitly given; i.e. anger, fear, happiness, joy, etc. The complete list of markers is given by the presence of punctuation marks (and sequences of them), words in capitals, emoticons, and quotation marks.

The second feature consists of a set of words that are often used by internet users as a subcode to communicate their messages more accurately. Words and abbreviations such as *lol* (Laughing out loud), *2U2* (To you too), *TGFF* (Thank God for Friday) are examples of such subcode. In order to obtain a reliable set of words to represent this feature, we used a common chat slang dictionary extracted from web [5] as a simple list of terms.

The following feature, context, is intended to identify common elements across the different classes of the corpus. To this end, we employed a cluster algorithm described in [1]. The result is a set of descriptive and discriminating words to represent each class. Such words are then used as descriptors of general contexts concerning both genre and age.

Emotionality is a feature to integrate information related to the communication of subjective matters through the selection of particular words. The Dictionary of Affect in Language ([5]) was used to represent this type of information. It is divided in three categories: Activation, Imagery, and Pleasantness. Each is intended to quantify the emotional content of words in terms of scores obtained from human raters.

The last feature is used to measure the semantic relatedness of the words. This is done in order to determine a threshold of semantic similarity among the different types of discourses profiled by the authors. The WordNet::Similarity toolkit described in [3] was used for obtaining the similarity.

In addition to the features above described, a list of Bag of Words (the most frequent words in the corpus) was used. Finally, the Jaccard similarity coefficient ([2]) was applied over the texts in order to focus on informative words rather than only on frequent ones.

[5] `http://www.chatslang.com/terms/common`

3 Experiment and Results

For our experiments we use a subset of 12.000 conversations from the *PAN 13 Training Corpus for Author Profiling Task*[6]. 9.000 conversations was used for training (1.500 each class) and 3.000 for test (500 each class). All the six classes included in the PAN 2013 training corpus were considered (female 10s - 30s and male 10s - 30s).

Each conversation is represented as a numerical vector in which each entry represent a feature. Then we make different combinations of the features proposed and we classified the conversations using various learning algorithms.

The first combination is called *SBF, Semantic-bases features* consist of $Semantic\ similarity + Signatures + ChatSlang + Emotionality$ measures. The second combination is composed of the *SBF + BOW*. The third combination is composed for *SBF + Jaccard Distance*.The fourth combination consist of *SBF + BOW + Jaccard Distance*. The fifth combination is only the *Jaccard Distance*. Sixth combination is *Jaccard Distance + BOW*, and the seventh consists of *Jaccard Distance + Context*. The learning algorithms applied to this combinations were Naive Bayes (NB) , Support Vector Machines (SVM), Multilayer Perceptron (MP), Decision tree (J48), and a bagging of classifiers (NB + SVM + J48).

The Table 1 introduces the results obtained from different experiments in terms of accuracy.

Experiments	NB	SVM	MP	J48	Bagging	Average
SBF	16.66	19.66	18.3	17.66	18.67	17.99
SBF+BOW	22	15.66	-	21.66	20.67	19.77
SBF+Jaccard Distance	19	20.60	18.33	15.33	20	18.31
SBF+Jaccard Distance+BOW	23	15.66	-	22	21	**20.22**
Jaccard Distance	22.33	17.33	22.33	14.67	17.66	18.11
Jaccard Distance+BOW	23	15.33	-	19.33	21.33	19.22
Jaccard Distance+Context	17.66	21.6	-	17.33	21.33	18.86

Table 1. Results of Author Profiling classifiers

In the MP column there are some results in blank because the dimensionality of numerical vectors for each conversation is very large; therefore, the algorithm did not converge with this experiments.

After analyzing the results above described, we removed some features in order to reduce the dimension of the vector. The ***SBF*** experiment was modified taking into account only the semantic similarity and emotionality measures ($\boldsymbol{SBF}_{\mathrm{M}}$). Finally, we carried out a new experiment using $\boldsymbol{SBF}_{\mathrm{M}}$***+Jaccard Distance+BOW***. This experiment was assessed with the **NB** classifier. The accuracy rate reported increased up to 23.66%.

[6] `http://pan.webis.de`

3.1 PAN Results

From the previous insights, we defined a final model which was integrated with the features: ***SBF***$_{\mathrm{M}}$***+Jaccard Distance+BOW***. According to our best results, the **NB** classifier was used to participate in the author profiling task in PAN 2013 competition. The results, considering both English and Spanish, are shown in the Table 2.

Task	**Accuracy**		
	total	Gender	Age
English	0.2816	0.5671	0.5061
Spanish	0.1757	0.4982	0.3554

Table 2. Author Profiling Evaluation PAN 2013

4 Conclusions

In this article we present an approach to identify author profiling. This research is based on semantic features. The set of features we use are: signatures, chat slang, context, emotionality, semantic similarity, Jaccard similarity and BOW. A tree structure was developed in order to weigh the set of features and select the best ones. After analyzing the results, we could realize that the author profiling task has a high level of overlap between classes; hence, the difficulty of correctly identifying the classes increases substantially. The future work consists of developing an algorithm for principal components analysis (PCA) in order to obtain highly discriminating features.

References

1. Karypis, G.: Cluto. A clustering toolkit. technical report 02-017. Tech. rep., University of Minnesota, Department of Computer Science. (2003)
2. Manning, C., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, MA, USA (1999)
3. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet::similarity - measuring the relatedness of concepts. In: Proceeding of the 9th National Conference on Artificial Intelligence (AAAI-04). pp. 1024–1025. Association for Computational Linguistics, Morristown, NJ, USA (2004)
4. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation (2012), DOI: 10.1007/s10579-012-9196-x
5. Whissell, C.: Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychological Reports 105(2), 509–521 (2009)