Author Profiling, instance-based Similarity Classification

Notebook for PAN at CLEF 2017

Yaritza Adame-Arcia[1], Daniel Castro-Castro[1], Reynier Ortega Bueno[1], Rafael Muñoz[2]

[1]Desarrollo de Aplicaciones, Tecnología y Sistemas DATYS, Cuba

yaritza.adame@datys.cu, {reynier.ortega, daniel.castro}@cerpamid.co.cu

[2]Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, España

rafael@dlsi.ua.es

Abstract. In digital documents analysis for forensic applications, when anonymous documents are presented and it is not possible with the available tools to determine the true author of the document, there are of vital importance methods that identify the characteristics of the Author Profile (Gender, Age, Personality, etc.). We propose to use a simple method of classification based on the similarity between objects, considering different features for documents representation: (a document corresponds to a set of tweets of a user), the terms used in the tweets, as well as characteristics of opinion and subjectivity presented in them. Our goal will be to classify, based on the content of the tweets, the Gender and language variety of an author from an unknown set of tweets corresponding to him. In the experiments we observed good results in Gender classification, but low values in language variety classification. We processed only the English dataset.

Keywords: Author profiling, instance-based classification, tweets gender classify, tweets language variety classify

1 Introduction

The PAN Profiling task for this edition is as follows: "Gender and language variety identification in Twitter. Demographics traits such as gender and language have so far investigated separately. In this task we will have participants with a corpus annotated with authors' gender and their specific variation of their native language:

- English (Australia, Canada, Great Britain, Ireland, New Zealand, United States)
- Spanish (Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela)
- Portuguese (Brazil, Portugal)
- Arabic (Egypt, Gulf, Levantine, Maghrebi)

Although we suggest to participate in both subtasks (gender and language identification) and in all languages, it is possible participating only in one of them and in some of the languages."

The proposal to identify these demographic traits in tweets implies that the natural language processing tools widely used for long documents analysis must be adapted to the features of the textual genre and the writing characteristics presented in tweets. We must emphasize that the complexity lies in the fact that for this genre there are no linguistic rules or writing standards. Language is informal, usually direct and full of emotions.

In past tasks of demographic traits identification on PAN evaluation framework [5] [6], tweet genre was used and many works presented used lexical content (words, informal text, jargon) and characteristic features of the genre (URL, hashtags, mentions, retweet, emoticons, etc.). The generality of the proposals uses the classic Bag of Words representation of documents, employing in addition to the mentioned features, n-grams of some of them, for example, words n-grams, lemmas n-grams, POS-Tagging (Part of Speech Grammatical Categories) n-grams, etc. The fundamental difference of the proposal of this year to previous proposals, lies in evaluating and classifying by variety of the language.

For the classification process, decision tree-based approximations have been used, as well as SVM by a large number of competitors and a few others have used distance-based approximations to predict the closest class [14][15].

We are interested in implementing a distance-based classification strategy and with this, use previous results presented in the Author Identification edition of 2015 [3]. We will combine features of the lexical content of the tweets, their characteristic features, and polarity and emotion features of previous works of our group used in tasks of sentence polarity classification. We will experimentally evaluate the differences between an instance-based proposal and a prototype-based proposal, in the same distance-based strategy.

2 Implemented methods

We used two classification strategies, considering two documents representation variants. An instance based representation of the documents, where the set of tweets of an author (for each author it is available her/his gender and language variety) represents a document and with this idea, for each class (female class, male class) we have a set of documents. The second variant is a prototype-based representation, where a single document is formed for each class, and this document is constructed with all the tweets of each of the sample authors per class.

Figure 1 shows graphically the architecture of our proposal with the instance-based strategy.

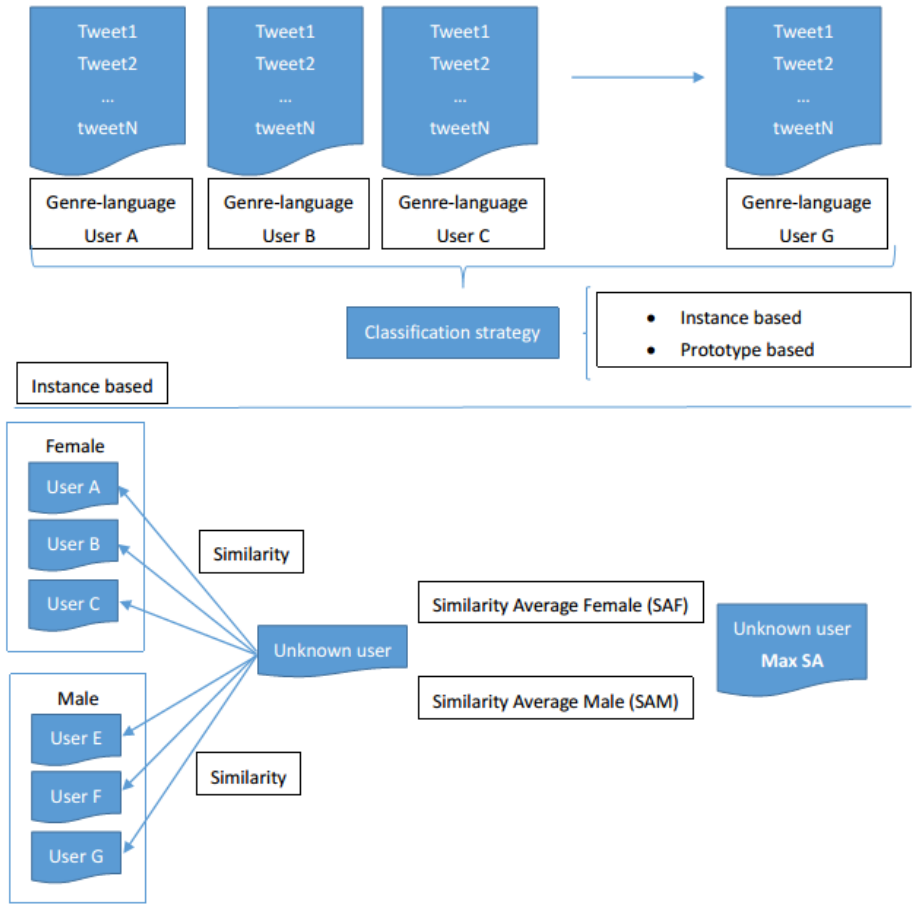

Fig. 1. Architecture proposed for author profiling. Instance based classification strategy.

2.1 Features and tweets pre-processing stage:

The first step correspond to build the documents that will be used as objects for the similarity calculation in the classification method. For each author, we receive the set of tweets that she/he wrote, and with the concatenation of these tweets is formed a document for this author. Remember that, of each author, what we have is the gender and the language variety. We perform a pre-processing of the document in two stages. In a first stage, we segment the tweets with a tokenizer offered in FreeLing [13] [http://nlp.lsi.upc.edu/freeling/], specialized for the processing of tweets. Subsequently we proceed to the expansion of short terms used and contractions, and characteristics traits that are used in tweets such as the Hashtags, URLs, mentions, are replaced by certain fixed patterns, those traits we consider the content does not contribute to differentiate between tweets of different profiles. After these transformations, we have nor-

malized the tweets a bit and next proceed to perform a syntactic analysis with the traditional POS-Tagging tools for English and Spanish according to the language of the tweets.

For the representation we use the classic Bag of Words and in this we integrate:

- The lexical terms, the lemmas of these and the grammatical category.
- Characteristics features of the tweets.
- Features of subjectivity and opinion mining analysis [7].

With the lexical terms and lemmas, we hope to differentiate the documents of each class, because some of this features are proper of their class. For example, for language variety, some terms are used by Colombians unlike the rest, and thus similarly for each variant of Spanish. Considering the frequency of use of grammatical categories, would allow us to differentiate between tweets written by the male gender and those written by the female gender. For example, in [17] it is exposed several differences in the use of words and different Part of Speech analyzing women and men writing style.

The characteristic features of the tweets we extract correspond to hashtags, the mention of author, the mark of retweet, the use of URL, the use of intensifications (capital letters, deformation of words by repetition of characters, use of admiration signs), use of laughter expressions, use of emoticons and the use of informal language. For each of these traits we consider the position in which they are used, that is, the number of times used at the beginning of the tweet, at the end or elsewhere.

Additionally, we include the analysis of the frequency of features with subjective information, for example, the number of positive or negative emoticons; the words used were categorized as Positive (P), High Positive (HP), Negative (N) and High Negative (HN), using the frequencies of this categories. We used a word polarity resource in Spanish and English taken from [12], resources of emotion in Spanish [8] [11] and for English [2], and finally the resources of appraisal for Spanish [9] [10] and English [1].

2.2 Classification stage:

For the classification of the set of tweets of an author in the Demographic traits of gender and language variety, we tried with two strategies. A strategy in which each document (set of author tweets) is used as an instance of the class to which it belongs and for the second strategy we construct a prototype of each class using the extracted features of the set of documents belonging to the class. Each of these strategies were evaluated with the tweets collections of the training set and was selected for the final evaluation, the one that showed more stable results in different executions.

In the instance-based strategy, it is calculated the similarity of the new document with each sample document of the class, and then is computed the average similarity obtained with the class. This analysis is done with each class of a Demographic Trait and the object is going to belong to the class with which it obtains greater average similarity. In the prototype-based strategy, the similarity of the new document is calculated with the class prototype. This analysis is done for each class and the object is going to belong to the class in which the similarity obtained was the highest (1-NN [4][16]).

The classification is done independently for each Author Demographic Trait, Gender classes (2 classes) and language variety (for English 6 classes and for Spanish 7 classes). Finally, the result is the combination of these two classifications.

3 Experiments and results

The initial experiments were performed with the training collection released for this year's 2017 task. We evaluated the accuracy obtained by performing a 2-cross fold validation. In addition, we considered the training collection of the 2015 edition for the Gender and Age classes. The description of these collections can be reviewed in [18] [6]. In Table 1 we include the values obtained in the tests with the two representation strategies, instance-based and prototype centroid-based one, using the collection of 2017. In table 2, we present the results with the collection of 2015.

Table 1. Accuracy in 2-cross fold validation train 2017

		Spanish	*English*
Instance based	gender	**0,6**	0,56
	Language variety	0,2	0,23
	join	0,12	0,14
Prototype based	gender	**0,63**	**0,65**
	Language variety	0,3	0,3
	join	0,19	0,2

Table 2. Accuracy in 2-cross fold validation train 2015

		Spanish	*English*
Instance based	gender	**0,68**	0,56
	Age	**0,46**	0,45
	join	0,29	0,21
Prototype based	gender	0,68	0,58
	Age	0,21	0,1
	join	0,17	0,1

Evaluating the results shown in the two tables, we consider that the classification is more stable with the instance-based strategy, so we decided to include this configuration in the evaluation of the task of this year. The results obtained can be observed in the summary published by the organizers and in the following table.

The results with the test dataset are shown in [18] and presented on the PAN web site. We got the lowest values of all the participants, and only run successfully for the English dataset. Comparing the results obtained using the BOW-baseline that uses the 1000 most frequent terms, we conclude that one of our problems is that we need to analyze and reduce the features used. We processed the dataset using the instance-based strategy and perhaps the results could be better if we used de prototype-based strategy whit feature selection methods.

4 Conclusions and future work

A representation that considers the terms used in tweets, is able to differentiate to a large extent the sets of tweets written by authors of different genres. The proposed subjectivity and opinion features allow improvements in classification, but they are not substantial improvements. In the evaluation we made with the collections of 2015, we verified that each of the sets of features separately allows good identifications of the genre and that their combination increases the values obtained. The classification in language variety maintains low results and to a great extent this is due to the little difference that is observed between some of these classes and that many terms used by the authors are of universal character and are standardized in the community.

We achieved the lowest values of all the team and considering that a baseline method using the 1000 most frequent terms in a Bag of Word representation got better results, then we need to do an exhaustive evaluation of our method.

We must work on features selection strategies and the analysis of representative objects to each of the classes. We propose to evaluate a classification with rejection or abstention for those users whose tweets do not contain characteristic features with their class, for example for the idea of language and not penalize so much the possible bad classifications.

5 References

1. Bloom, K., Garg, N., & Argamon, S. Extracting Appraisal Expressions. In Proceedings of NAACL HLT 2007. Rochester, NY: Association for Computational Linguistics. pp. 308–315. 2007
2. Carlos Strapparava, Valitutti Ro. WordNet-Affect: an Affective Extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation. 2004. 1083--1086
3. Daniel Castro, Yaritza Adame, María Peláez Brioso, Rafael Muñoz: Authorship Verification, combining Linguistic Features and Different Similarity Functions. CLEF (Working Notes) 2015
4. Efstathios Stamatatos. A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, Volume 60, Issue 3, pages 538-556, March 2009.
5. Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, Benno Stein: Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. CLEF (Working Notes) 2016: 750-784
6. Francisco M. Rangel Pardo, Fabio Celli, Paolo Rosso, Martin Potthast, Benno Stein, Walter Daelemans: Overview of the 3rd Author Profiling Task at PAN 2015. CLEF (Working Notes) 2015
7. Francisco Rangel, Paolo Rosso. On the Impact of Emotions on Author Profiling. In: Information Processing & Management, vol. 52, issue 1, pp. 73-92
8. Grigori Sidorov, Sabino Miranda-Jiménez, Francisco Viveros-Jiménez, Alexander Gelbukh, Noé Castro-Sánchez, Francisco Velásquez, Ismael Díaz-Rangel, Sergio Suárez-Guerra, Alejandro Treviño, and Juan Gordon. Empirical Study of Opinion Mining in Spanish Tweets. LNAI 7629, 2012, pp. 1-14.

9. Hernández, L., López-Lopez, A., & Medina-Pagola, J. E. (2009). Recognizing Polarity and Attitude of Words in Text. In In Proc. F 14th Portuguese Conference on Artificial Intelligence, (EPIA'2009) (pp. 525–536). Aveiro, Portugal.
10. Hernández, L., López-Lopez, A., & Pagola, J. E. M. (2011). Classification of Attitude Words for Opinions Mining. International Journal of Computational Linguistics and Applications, 2(1–2), 267–283.
11. Ismael Díaz Rangel, Grigori Sidorov, Sergio Suárez-Guerra. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomazein , 29, 23 p., 2014, DOI 10.7764/onomazein.29.5
12. Jose Manuel Yero Moreno, Reynier Ortega Bueno. Método no supervisado para la clasificación de polaridad en Twitter. VII Conferencia Internacional de Ingeniería Eléctrica. . pp. 1 - 4. Jun, 2014. ISBN: 978-959-207-529-0.
13. Lluís Padró, Evgeny Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.
14. Mirco Kocher, Jacques Savoy: UniNE at CLEF 2016: Author Profiling. CLEF (Working Notes) 2016: 903-911
15. Maria José Garciarena Ucelay, Maria Paula Villegas, Dario G. Funez, Leticia C. Cagnina, Marcelo Luis Errecalde, Gabriela Ramírez-de-la-Rosa, Esaú Villatoro-Tello: Profile-based Approach for Age and Gender Identification. CLEF (Working Notes) 2016: 864-873
16. Patrick Juola. Authorship Attribution. In Foundations and Trends in Information Retrieval, Volume 1, Issue 3, March 2008.
17. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates 71 (2001)
18. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)