CerpamidUA at MexA3T 2019: Transition Point Proposal

Daniel Castro Castro[1][0000−0001−9102−7601], María Fernanda Artigas Herold[2],
Reynier Ortega Bueno[1], and Rafael Muñoz[3]

[1] Center for Pattern Recognition and Data Mining, Cuba
{daniel.castro, reynier.ortega}@cerpamid.co.cu
http://www.cerpamid.co.cu
[2] Oriente University, Cuba
maria.artigas@estudiantes.uo.edu.cu
[3] Department of Software and Computing systems, Alicante University, Spain
rafael@dlsi.ua.es
http://www.dlsi.ua.es

Abstract. Author Profiling is an important field for detection of demographic characteristics of users based on texts written by him. Our main contribution is focused in determining a reduced subset of features that represent frequent lexical words for each profile of Mexican twitters. The new subset of features was obtained considering the frequency of words in a profile (e.g.: students), employing the theory of Transition Points. All the objects are represented in this new feature space conformed by all the reduced subset computed for each class or profile. The classification phase was carried out using Support Vector Machines provided by the Weka platform. The results obtained were good for Gender, but needs more efforts for Location and Occupation, because, the main factor that affects the results correspond to scenarios with unbalanced class distribution that impact the construction of the reduced vocabulary.

Keywords: Author Profiling · Transition Point · Mexican Twitter Profiling.

1 Introduction

The modern society is characterized by an impressive use of digital technology and in particular to socialize using Social Network platforms in which emotions, ideas, new information, etc, are expressed. Users share their information using image, text, videos and other resources. All the available public information of an user, and in particular text and image, could be used to determine demographic attributes of him, such as, gender, age, personality, level of scholarship and others, and this is the key question in study in the field of Author Profiling

(AP) analysis.
In 2018, it was proposed the MexA3T task for Author Profiling and Aggressiveness analysis focused on Mexican tweets [3]. The AP task comprises the detection of Place of Residence and Occupation of an user profile based on the set of tweets written by him. As it was exposed in the overview [3], it was a challenging task and for that reason they relaunch a similar task; including the analysis of Gender characteristics.
An important difference of this year [1] with respect to the previous task is that an user profile is distributed not only using the text of the tweets, but also images were incorporated on the profiles. This will allow the use of Text and Image for profiling classification and it is not necessary to use both information.
The principal evaluation Forum for Authorship Analysis over several years has been the PAN Lab at CLEF and in particular it has evaluated the AP [5] task considering the identification of Gender, Personality, Age, etc.
In MexA3T 2018 AP task, participated 4 teams [9] [2] [6] [8], the majority of them used an approach based on SVM classification and representation of text employing as features n-grams of character and lexical tokens. The MXAA [9] team was in average the top ranked and it used a feature selection and term weighting strategies that allowed them to achieve very good results.

2 Proposal for MexA3T 2019

Our main contribution is focused in determining a reduced subset of features that represent frequent lexical words for each profile of Mexican tweets writers. The new subset of features was obtained considering the frequency of words in a profile (e.g.: students), by using of, the theory of Transition Point [7]. All the objects are represented in this new feature space conformed by all the reduced subset in each class or profile. The classification phase was performed using Support Vector Machines provided by the Weka [4] platform with default configuration.

2.1 Transition Point

The architecture for the dimensionality reduction of the vocabulary based on Transition Point Method is illustrated in the Figure 1.

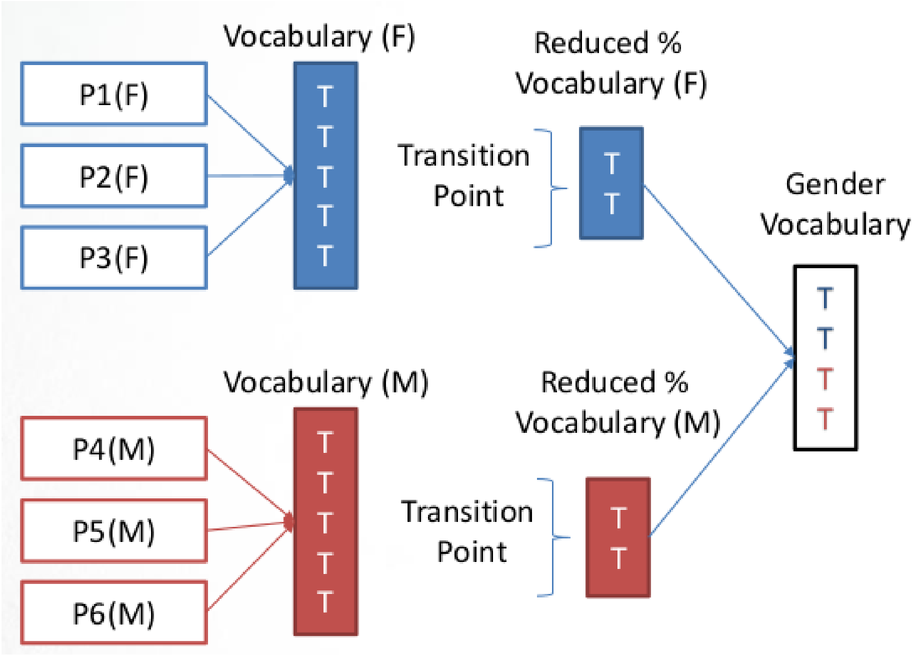

Fig. 1. Architecture for dimensionality reduction using Transition Point

Transition Point (TP), refers to a frequency value in the vocabulary that delimit a frontier in which the terms of the vocabulary are relevant to the class and with high presence in objects of that class. It is based on the fundamentals studied and proposed by [11], who formulated the Law of word frequencies in a text, Zipf's Law. We first build a vocabulary for each profile (e.g., a vocabulary for male profile and a vocabulary for female profile) and each term of the vocabulary is associated with the frequency of occurrence in the tweets of its correspondence profile. The TP is calculated for each vocabulary profile (V_p) and using this, it is selected a percentage of tokens with frequency close to the value of TP. The new vocabulary for a profile class (Gender Profile) is formed by the union of the tokens present in the reduced vocabulary obtained for each profile.

2.2 Tweet representation

The profiles are conformed by several tweets written by users. We consider a tweet as a document and represent the tweet by the tokens extracted using a Natural Language Processing Tools (NLPt). We used the FreeLing [10] NLPt and executed a first representation based on the tokens extracted by the tokenizer. A second representation was built considering the lemmas of the tokens. In each of these representations, the features are weighted by a normalized frequency of occurrence.

2.3 Machine Learning Method

The supervised classification phase is done using SVM implemented in Weka platform with the default parameters. An user profile is conformed by all the tweets written by him, and afterwards each tweet is represented in the new reduced vocabulary, it is conformed a prototype formed by a centroid of all the tweets.

3 Evaluation, Results and Discussion

The dataset distributed contains profiles for three classes: Gender, Location and Occupation [1] and the difference with respect to MexA3T 2018 task is the Gender class. Particularly, the Gender dataset is balanced for each class, female and male, but the Location and Occupation dataset is unbalanced.
The evaluation was made using F-measure by class, accuracy and F-average in a profile.
The row CerpamidUA-Gender-Text-run1 used as vocabulary the extraction of 1 percent of tokens from the vocabulary of each class and the representation based on words extracted by a tokenizer. The row CerpamidUA-Gender-Text-run2 considered 10 percent of tokens and the representation based on lemmas.
In Table 1, is illustrated the results obtained for gender classification.

Table 1. Gender results.

Team	F(P,R)	Acc	P	R
CerpamidUA-Gender-Text-run2	**0.83**	0.83	0.84	0.83
CerpamidUA-Gender-Text-run1	0.83	0.83	0.83	0.83
CIC-VCR-Secondary-Gender-Image	0.52	0.52	0.52	0.52
CIC-VCR-Gender-Image	0.47	0.48	0.48	0.48

The results obtained by run2 are similar than those of run1. In general the results are good, due to the balanced scenarios in both classes male and female. It is also important to notice that the representation based on lemma has less dimension than the representation based on tokens and the proposal to obtain a new vocabulary considering the TP, reduced the dimension dramaticaly obtaining good results.
In Table 2, is illustrated the result obtained for Location classification.

Table 2. Location results.

Team	F(P,R)	Acc	center	southeast	northwest	north	northeast	west
CerpamidUA-Location-run2	0.50	**0.63**	**0.68**	0.38	**0.66**	0.16	**0.70**	0.28
CerpamidUA-Location-run1	0.48	**0.61**	**0.70**	0.39	**0.66**	0.25	**0.72**	0.26
CIC-VCR-Secondary-Gender-Image	0.14	0.23	0.41	0.04	0.09	0.02	0.20	0.10
CIC-VCR-Gender-Image	0.10	0.16	0.36	0.00	0.08	0.04	0.13	0.02

The results for Location classification are not high. The results are modest , we suppose that this drop, can be caused by the unbalance of the datasets. The majority classes get the best results, but the classes with few profiles achieved worse values. The accuracy values reflect that the majority class classifies very good its objects. The main problem is related to the vocabulary constructed, because the class with few objects contributes less with new tokens corresponding to it.
In Table 3, is illustrated the results obtained for Occupation classification, and the analysis of the results reflects similar conclusions than those explained for Location classification.

Table 3. Occupation results.

Team	F(P,R)	Acc	others	arts	student	social	sciences	sports	admin	health
CerpamidUA-Occupation-run2	0.39	**0.65**	0.10	0.25	**0.85**	**0.56**	0.19	0.30	0.51	0.25
CerpamidUA-Occupation-run1	0.38	**0.66**	0.13	0.33	**0.86**	**0.55**	0.20	0.35	0.47	0.24
CIC-VCR-Secondary-Gender-Image	0.11	0.26	0.00	0.09	0.44	0.13	0.06	0.00	0.21	0.00
CIC-VCR-Gender-Image	0.09	0.23	0.00	0.11	0.43	0.07	0.04	0.00	0.09	0.00

4 Conclusion and Future Work

In class with few document the results were low, determined by the scarce variety of the words of these classes in the vocabulary generated using TP. It was obtained very good results in the identification of gender, conditioned by the balance between classes. The weight of the features should be evaluated considering the difference between dictionaries per class and the importance of each word in the new reduced vocabulary.

References

1. Aragón, M.E., Álvarez-Carmona, M.Á., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Moctezuma, D.: Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, September (2019)
2. Aragón, M.E., López-Monroy, A.P.: Author profiling and aggressiveness detection in spanish tweets: Mex-a3t 2018. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 134–139 (2018), http://ceur-ws.org/Vol-2150/MEX-A3T_paper7.pdf
3. Ángel Álvarez Carmona, M., Guzmán-Falcón, E., y Gómez, M.M., Escalante, H.J., nor Pineda, L.V., Reyes-Meza, V., Sulayes, A.R.: Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the

Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 74–96 (2018), http://ceur-ws.org/Vol-2150/overview-mex-a3t.pdf

4. Eibe Frank, M.A.H., Witten, I.H.: The weka workbench. online appendix for "data mining: Practical machine learning tools and techniques" (2016)
5. Francisco Manuel, R.P., y Gómez, M.M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: Cross-domain authorship attribution and style change detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France. CEUR-WS.org (sep 2018), http://ceur-ws.org/Vol-2125/
6. Graff, M., Miranda-Jiménez, S., Tellez, E.S., Moctezuma, D., Salgado, V., Ortiz-Bejar, J., Sánchez, C.N.: Ingeotec at mex-a3t: Author profiling and aggressiveness analysis in twitter using μtc and evomsa. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 128–133 (2018), http://ceur-ws.org/Vol-2150/MEX-A3T_paper6.pdf
7. Jiménez-Salazar, H., Pinto, D., Rosso, P.: Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos. Procesamiento del Lenguaje Natural **35** (2005), http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2991/1485
8. Markov, I., Gómez-Adorno, H., Rosales, M.J., Sidorov, G.: Cic-gil approach to author profiling in spanish tweets: Location and occupation. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 97–101 (2018), http://ceur-ws.org/Vol-2150/MEX-A3T_paper1.pdf
9. Ortega-Mendoza, R.M., López-Monroy, A.P.: The winning approach for author profiling of mexican users in twitter at mex.a3t@ibereval-2018. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. pp. 140–148 (2018), http://ceur-ws.org/Vol-2150/MEX-A3T_paper8.pdf
10. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012. pp. 2473–2479 (2012), http://www.lrec-conf.org/proceedings/lrec2012/summaries/430.html
11. Zipf, G.K.: Human behaviour and the principle of least effort. Addison-Wesley (1949)