CIC-GIL Approach to Author Profiling in Spanish Tweets: Location and Occupation

CIC-GIL Approach to Author Profiling in Spanish Tweets: Location and Occupation

Ilia Markov[1], Helena Gómez-Adorno[2], Mónica Jasso-Rosales[2], and Grigori Sidorov[1]

[1] Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico City, Mexico
imarkov@nlp.cic.ipn.mx, sidorov@cic.ipn.mx,
[2] Universidad Nacional Autónoma de México (UNAM), Engineering Institute (II), Mexico City, Mexico
hgomeza@iingen.unam.mx, mjassor@iingen.unam.mx

Abstract. We present the CIC-GIL approach to the author profiling (AP) task at MEX-A3T 2018. The task consists of two subtasks: identification of authors' location (6-way) and occupation (8-way) in a corpus of Mexican Spanish tweets. We used the logistic regression algorithm trained on typed character n-grams, function-word n-grams, and regionalisms for location identification, and typed character n-grams with several modifications for occupation identification. Our best run showed F1-macro score of 73.63% for location and 48.94% for occupation identification. The results are competitive with other participating teams; in particular, our best run was ranked fourth in the shared task.

Keywords: author profiling, location identification, occupation identification, social media, n-grams, Spanish, machine learning

1 Introduction

Author profiling (AP) is the task of identifying the author's demographics, such as age, gender, personality traits, native language, place of residence, and occupation based on a sample of his or her writing. AP is useful for a variety of purposes, including security, marketing, and forensic applications.

The interest in the AP task is maintained through the annual organization of the PAN evaluation campaign[3] – one of the main *fora* regarding tasks related with authorship analysis. The author profiling task at MEX-A3T [1] focuses on Mexican Spanish and covers two aspects of author profiling, which have not been previously explored in any related competitions: place of residence (location) and occupation of the author.

We approach the task from a machine-learning perspective, as a multi-class classification problem. Further, we briefly describe the dataset used in the competition, and then focus on the applied pre-processing steps, features, and the configuration of our system.

[3] http://pan.webis.de

2 Data

The training corpus provided by the organizers consists of 3,470 documents[4], and is quite imbalanced in terms of both location and occupation. The corpus statistics is provided in Table 1. It shows the number of documents (No. of docs) and the percentage (%) for each class, as well as the average (Avg.), minimum (Min.), and maximum (Max.) document length measured in terms of characters (after applying pre-processing steps described below). A more detailed description of the corpus can be found in [1].

Table 1. Training corpus statistics.

Class	No. of docs (%)	Avg. length	Min. length	Max. length
Location				
Center	1,252 (36%)	114,088	15	583,086
Northeast	911 (26%)	60,163	25	372,015
Northwest	575 (17%)	53,580	68	268,186
West	314 (9%)	114,127	72	544,112
Southeast	312 (9%)	106,647	995	363,068
North	106 (3%)	92,592	999	310,205
Occupation				
Student	1,637 (47%)	51,889	15	296,659
Administrative	624 (18%)	128,477	109	360,812
Social	563 (16%)	121,248	121	583,086
Arts	239 (8%)	116,794	346	301,275
Sciences	182 (5%)	115,148	560	321,545
Health	105 (3%)	103,072	495	310,205
Others	75 (2%)	131,580	999	359,889
Sports	45 (1%)	98,753	999	197,539

3 Methodology

3.1 Pre-processing steps

Pre-processing has proved to be a useful strategy for the AP task [2]. We applied several pre-processing steps in order to aid typed character n-gram features to capture relevant information: (i) We performed lowercasing. (ii) We replaced all digits by the same symbol (e.g., 2,123 → 0,000), since we are not interested in the actual number, while the frequency of digits and their length may provide useful information to the classifier. (iii) We replaced user mentions (@user), user hashtag mentions (#tag), picture links, and URL mentions by different symbols (@user → 1, #tag → 2, picture link → 3, URL → 4) in order to keep information about their occurrence and remove information about the exact user/tag/link/URL. For location, we reduced user mentions, hashtag mentions, picture links, and URL mentions to the same symbol (@user → 1, #tag → 1, picture link → 1, URL → 1). (iv) We replaced slang words by their standardized version from the Spanish social media lexicon, as proposed in [2].

[4] We removed 30 empty documents from the dataset.

3.2 Features

Typed character n-grams Typed character n-grams, i.e., character n-grams classified into 10 categories based on affixes, words, and punctuation [3] have proved to be indicative features for other subtasks of author profiling, e.g., native language identification [4], identification of gender and language variety, including when different varieties of Spanish are concerned [5, 6].

For location identification, we used all categories of typed character n-grams ($n = 4$). For occupation, we conducted an ablation study in order to identify the most indicative typed character n-gram categories. We found that the *middle-punctuation* n-grams, which capture the frequency of punctuation marks, did not contribute to the result, and therefore were excluded. Additional weight, on the contrary, was assigned to the *middle-word* n-gram features (the most indicative category based on the ablation study), that is, we triplicated *middle-word* n-gram features: we used three different features, e.g., 'eatu-1', 'eatu-2', 'eatu-3', instead of one feature 'eatu'.

Function-word n-grams Function words (FWs) belong to a set of closed-class words and represent relations rather than propositional content. Examples of function words include articles, prepositions, determiners, conjunctions, and auxiliary verbs. FW n-grams are composed of n consecutive FWs omitting all the tokens in between.

We used the set of 313 Spanish function words from the Natural Language Toolkit (NLTK)[5] and built FW n-grams ($n = 2$). FW n-gram features were used only for location identification.

Regionalisms We developed a lexicon of regionalisms (words commonly used in a particular geographic area) used in three regions of Mexico: North, Center, and South. The lexicon contains 614 regionalisms obtained from the following websites:[6]

- `http://lexiquetos.org/chilanguismos/`
- `http://www.multimedios.com/telediario/tendencias/diccionario-basico-regio.html`
- `http://www.chicaregia.com/2006/06/diccionario-vocabulario-regio/`
- `http://www.sobrino.net/Dzidzantun/d_yuc.htm`

We used the regionalisms from the lexicon as features for location identification.

3.3 Experimental setup

Classifier We used the scikit-learn [7] implementation of the logistic regression (LR) algorithm, which showed higher results than other machine-learning algorithms we examined: SVM and multinomial Naive Bayes.

[5] http://www.nltk.org
[6] The lexicon is available upon request.

Weighting We used term frequency (tf) weighting scheme, i.e., the number of times a term occurs in a document. Other weighting schemes we examined – binary, tf–idf, and log-entropy – deteriorated our results.

Threshold In our primary run, we considered only those features that occur in at least 15 documents in the entire corpus and that occur in at least two documents in the corpus. In our secondary run, we did not use any threshold and considered only those features that occur in at least two documents in the corpus.

Evaluation For the evaluation of our system, we conducted experiments under 5-fold cross-validation on the training corpus measuring the results in terms of F1-macro score (the official metric).

4 Results

The 5-fold cross-validation results (F1-macro score, %) for our two runs on the training data, as well as the official results on the test set for our team and the best results in the shared task are shown in Table 2. The number of features (No.) for each run, as well as the majority and bag-of-words (BoW) baselines are also provided.

Our secondary run (without threshold) showed higher results on the test set for both location and occupation: 73.63% and 48.94%, respectively. The obtained results are much higher than the baselines.

Concerning the contribution of the regionalisms to location identification, in our experiments on the training data these features contributed about 1.5% to the overall 5-fold cross-validation F1-score.

Table 2. 5-fold cross-validation results on the training corpus (Train) and the official results on the test set (Test) in terms of F1-macro (%) for identification of location and occupation. The best results in the competition and our highest results are in bold typeface.

	Location			**Occupation**			**Average**	
	Train	**Test**	**No.**	**Train**	**Test**	**No.**	**Train**	**Test**
Shared task best	–	**83.88**	–	–	**51.22**	–	–	**67.12**
Majority baseline	8.84	–	–	8.01	–	–	8.43	–
BoW baseline	57.85	–	2,180,163	35.96	–	2,180,163	46.91	–
Primary run	**72.90**	73.10	175,088	**45.46**	47.27	306,647	**59.18**	60.19
Secondary run	71.84	**73.63**	666,078	44.21	**48.94**	1,382,696	58.03	**61.29**

5 Conclusions

We presented the CIC-GIL approach for identification of location and occupation in a corpus composed of Twitter messages in Mexican Spanish. Our simple approach achieved higher results than the baseline: 73.63% F1-macro score for location and 48.94% for occupation on the test set, and was placed fourth in the shared task.

One of the directions for future work would be to use doc2vec document embeddings. This strategy has proved to be useful for the AP task [8]. We will also examine other approaches on this dataset, such as the statistical-based approach described in [9].

References

1. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. In: Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September. (2018)
2. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience **2016** (2016) 13 pages
3. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies, Denver, CO, USA, ACL (2015) 93–102
4. Markov, I., Chen, L., Strapparava, C., Sidorov, G.: CIC-FBK approach to native language identification. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, ACL (2017) 374–381
5. Markov, I., Gómez-Adorno, H., Sidorov, G.: Language- and subtask-dependent feature selection and classifier parameter tuning for author profiling. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. Volume 1866 of CEUR Workshop Proceedings., Dublin, Ireland, CLEF and CEUR-WS.org (2017)
6. Gómez-Adorno, H., Markov, I., Baptista, J., Sidorov, G., Pinto, D.: Discriminating between similar languages using a combination of typed and untyped character n-grams and words. In: Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects, ACL (2017) 137–145
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
8. Markov, I., Gómez-Adorno, H., Posadas-Durán, J., Sidorov, G., Gelbukh, A.: Author profiling with doc2vec neural network-based document embeddings. In: Proceedings of the 15th Mexican International Conference on Artificial Intelligence, MICAI 2016. Volume 10062., Cancún, Mexico, Part II, LNAI, Springer (2017) 117–131
9. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: The winning approach to cross-genre gender identification in Russian at RUSProfiling 2017. In: FIRE 2017 Working Notes. Volume 2036 of FIRE'17., Bangalore, India, CEUR-WS.org (2017) 20–24