Twitter Author Profiling Using Word Embeddings and Logistic Regression

Notebook for PAN at CLEF 2017

Liliya Akhtyamova[1], John Cardiff[1], Andrey Ignatov[2]

[1] Institute of Technology Tallaght, Ireland
liliya.akhtyamova@postgrad.ittdublin.ie,
john.cardiff@it-tallaght.ie

[2] ETH Zurich, Switzerland

Abstract The general goal of the author profiling task is to determine various social and demographic aspects of the author based on his pieces of writing. In this work, we propose an approach that combines word embeddings and classical logistic regression for identifying author gender and language variety based on the corresponding tweets. The model was trained on PAN 2017 Twitter Corpus that contains data for English, Spanish, Portuguese and Arabic languages from more than 11 thousand authors. Due to its simplicity, the proposed solution can be treated as a baseline for both gender and language variety identification subtasks.

1 Introduction

With the world becoming more digital, the personal data of internet users such as their gender, age, ethnicity or personality type is playing more and more important role in the modern life. This information can be used for providing relevant search results, recommending appropriate connections in social networks, fraud prediction or personalized advertising. While there have been already proposed numerous solutions to tackle this kind of problems, the majority of them were relying on hand-designed features and various heuristics. In this works, we propose a simple fully-automated way of performing author profiling based on text data. The detailed architecture of our system is described in the following sections.

2 Models and Methods

In this section, we give an overview of the proposed method and describe its main components.

2.1 Dataset

In this work, we use PAN 2017 Author Profiling Dataset [2] published along with the corresponding shared task [1]. This dataset contains tweets from 11400 users for English, Spanish, Portuguese and Arabic languages. For each tweet, there is an information

about his author id, author gender and author language variety. The language variations presented in this dataset are the following:

- *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

Overall, there are 360K, 420K, 120K and 240K tweets for English, Spanish, Portuguese and Arabic languages, respectively. These tweets are further used as an input data for our algorithm.

2.2 Input Processing

In this task, the input to our classification model has the form of the sentence (tweet) $\mathbf{T}$ that is treated as an ordered sequence of words: $\mathbf{T} = \{w_1, w_2, ..., w_N\}$. First, plain words are mapped to their vector representations using a pre-trained word embedding model, which in our case is word2vec. The resulting representations are summed and averaged to form a single sentence vector $\mathbf{M_T}$, which dimensionality d is equal to the dimensionality of word embeddings. This vector is then passed to Logistic Regression classification algorithm.

2.3 Logistic Regression

Multinomial Logistic Regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels. Thus it is an extension of logistic regression, which analyzes dichotomous (binary) dependents. Like all linear regressions, the multinomial regression is a predictive analysis. Multinomial regression is used to describe data and to explain the relationship between one dependent nominal variable and one or more continuous-level(interval or ratio scale) independent variables.

2.4 Alternatives

In our initial experiments we have additionally tried a number of different methods to solve the considered author profiling problem. Among them were bag-of-words model, that takes into account the multiplicity of the appearing words in each sentence, and CNN-based solution that performs sentence classification based on the sentence matrix [4]. Besides that, we have also tried various classifiers: Random Forest, Linear Regression, Naive Bayes, SVMs, etc. However, our experiments revealed that these solutions demonstrate the same or worse performance compared to the one proposed in this work, therefore it was chosen for our submission and final evaluation.

Table 1. The results of the proposed model on both subtasks for four languages.

Language	Language variety identification accuracy, %	Gender identification accuracy, %
English	58.13	74.46
Spanish	80.32	69.46
Portuguese	97.63	68.50
Arabic	44.88	64.25

3 Experiments

To generate word embeddings, we trained a CBOW word2vec model on the initial twitter corpus with a context window of size 5, and a vector dimensionality of 300. A separate word2vec model was trained for each major language (4 models in total), to do this Python *gensim* [3] library was used. *Numpy* library was utilized for general array manipulation, and the implementation of Logistic Regression was taken from *sklearn* machine learning library. The classifier was trained to minimize L_2 loss function, the regularization coefficient was set to $C = 0.1$.

The dataset was split into tow subsets: 90% of the data was used for training the model, and the rest 10% — for validation. The final results on the test dataset for each language are presented in the table 1. The proposed model was able to achieve the accuracy of 64% for Arabic language and 68 – 74% for the rest languages in the gender identification subtask. The results in the language identification task were highly dependent on the number of predictive classes and the difference between the dialects: 97% for Portuguese (2 classes), 44% for Arabic (4 classes), 58% for English (6 classes) and 80% for Spanish (7 classes). Relatively week results for Arabic language can be explained by the fact that the original hieroglyphs were encoded into unicode, and thus some relevance between similar hieroglyphs was lost.

4 Conclusion and Future Work

In this work, we presented an approach for gender and language variety identification on twitter data, that is based on the combination of word embeddings and logistic regression models. The proposed solution has the benefits of requiring no hand-designed features and being applicable to various nlp domains without a need for modifications in the implementation.

References

1. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017)

2. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs
3. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
4. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 959–962. ACM (2015)