UO_4to @ TAG-it 2020: Ensemble of Machine Learning Methods

Maria Fernanda Artigas Herold
Computer Science Department, Universidad de Oriente, Santiago de Cuba, Cuba
nanda.ah@nauta.cu

Daniel Castro Castro
Computer Science Department, Universidad de Oriente, Santiago de Cuba, Cuba
danielcc@uo.edu.cu

Abstract

This paper describes the proposal presented in the TAG-it author profiling task from EVALITA 2020 for sub-task 1. The main objective is to predict gender and age of some blog users by their posts, as well as topic they wrote about. Our proposal uses an ensemble of machine learning algorithms with three of the most used classifiers and language model of the n-grams of characters represented in a Bag of Word. To face this task we presented two different strategies aimed at finding the best possible results.

1 Introduction

With the growing development of technology and the frequent use of new forms of interactions and communications, Internet users spend more time sharing their ideas, thoughts, feelings and interests through social networks with diverse purposes, whether of personal businesses, self-expression, socialization, scientific, commercial, etc. In social media people often share their personal data, contact information, jobs, criteria and, in general, very useful information that can be used in research purposes about the behavior of people, development of marketing strategies and political campaigns, to serve various forensics applications, as well as strategies to determine certain demographic attributes of the person such as age, sex, characteristics of personality, geographic origins and even their occupation.

Precisely, one of the purposes of Natural Language Processing (NLP) research is to analyze the information obtained from users to create systems capable of extracting significant characteristics and improving the automatic understanding of written text.

Author Profiling (AP) is the main branch of NLP that studies the analysis of information to determine several demographic aspects of author such as age and gender given a set of documents presumably written by him, and recently some aspects such as the personality and occupation have also been included. The increased integration of social media in people's daily lives have made them a rich source of textual data for author profiling since data could be mined from the web, including emails and blogs, but there are still limitations in using social media as data source because data obtained may not always be reliable or accurate. Users used to provide false information about themselves that difficult the correct development of the task.

Document classification, also known as text tagging, is currently one of the most important subtask of Text Mining and NLP where the general idea is assign automatically one or more classes or categories in a set of predefined tags to a document using machine learning algorithms based on its content. Documents may be classified according to the subject, author or any other class that could be of interest in the research, as well as age and gender.

Recognized by the community, there is a theoretical evaluation framework, known as PAN[1], which encompasses authorship detection, author profiling, sentiment analysis, among others. On this platform, people can present and share their work, find out about the topics covered in previous works and participate in the tasks that are proposed each year for the community.

[1] https://pan.webis.de/

In 2019, at the PAN@CLEF evaluation forum (Rangel and Rosso, 2019), it was presented the Bots an Gender author profiling tasks, whose objective was determine if the author of a Twitter feed, in Spanish or English, had been written by a robot or a human, and in case of human, the gender should also be determined. To resolve this task, organizers proposed a set of baselines with models of n-grams of characters and words representation with a vocabulary reduction varying the parameters according to a few certain of configurations.

Another forum where the subject of author profiling has been worked on is MexA3T[2], another domain different from PAN for Spanish variants where generally works with the analysis of Mexican tweets. In 2019, it was proposed the MexA3T task for Author Profiling and Aggressiveness analysis focused on Mexican tweets (Aragón, 2019) as a follow-up of the task proposed in 2018 (Álvarez, 2018). The AP task comprises the detection of Place of Residence, Occupation and Gender of an user profile based on the set of tweets written by him. An user profile was distributed not only using the text of the tweets, but also images were incorporated on the profiles.

Several authors base their approaches on feature engineering and traditional machine learning classifiers. In previous works, methods have been proposed that work with comprising content-based (bag of words, word n-grams, term vectors, dictionary words), feature reduction (Castro, 2019) where the most used technique has been the selection of a subset of the most frequent features, stylistic-based features (frequencies, punctuation, POS, Twitter-specific elements, slang words) and approaches based on neural networks (CNN, LSTM) (Valdez, 2019).

2 TAG-it 2020

Despite the fact that Text Mining and NLP tasks focus a lot on the most used languages such as English and Spanish, others languages are also widely covered in several important forums. EVALITA[3] is a platform which promotes NLP tasks specifically for Italian language providing a shared framework where different systems and approaches can be evaluated in a consistent manner that has been working since 2007.

This year, TAG-it: Topic, Age and Gender Prediction for Italian from EVALITA (Cimino, 2020) propose three different sub-task of AP. The first one (subtask1) with the aim of predicting gender, age (in an age range, eg: 30-39) and the topic treated by the author given a collection of documents written by him/her in a blog, the three classes at once. The second one (subtask2a): for predicting gender only, and the third one (subtask2b): for predicting age.

For this task, a training corpus composed by texts written by users in a blog was offered, where each user has multiple posts. The information per user varies in length and quantity, in addition to the fact that the data is unbalanced for each class, which is not helpful for the training in classification task models.

2.1 Our method

According to the data corpus provided, our proposal is focused on classifying documents using a Bag of Word of n-grams characters representation, a feature reduction by a predefined number and an ensemble of machine learning algorithms: Random Forest, Support Vector Machine (SVM) and Centroid Nearest Neighbor classifiers, see Figure.1. We also consider Tf or a Tf-Idf as the weight of features.

We participate in the subtask1 where we present two different strategies. First we adjust the values of the parameters n for numbers of n-grams, k for feature reduction and the calculation of TF-IDF or not to the classification of each profile independently using a different configuration in each one according to the best results obtained in the individual classification. In the second proposal we adjust a general parameter and use the same configuration in the three profiles classification.

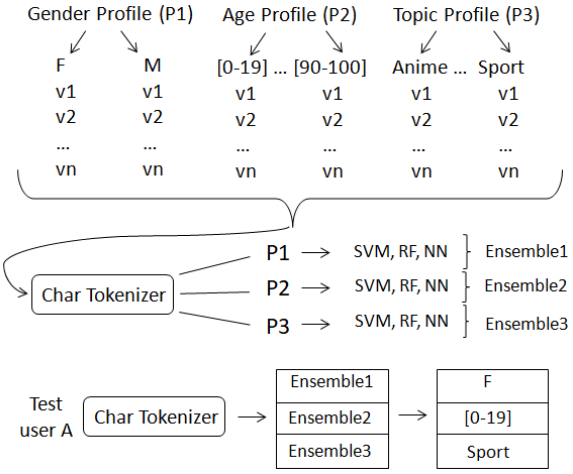

Figure.1 Ensemble architecture representation.

[2] https://sites.google.com/view/mex-a3t/
[3] http://www.evalita.it/

To represent the documents in a Bag of Word (BoW) model, we segment and preprocess the corpus and construct a vector of n-grams of characters ordered from highest to lowest by their respective frequency in the text per document. The parameters that we established for each configuration were: the n-grams of character representation, a size n from 1 to 5 characters and a number of 100, 500 and 1000 for feature reduction. Also for the weighing of the elements was considered the calculation of TF or TF-IDF, depending on the case, defined as follow:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{ij}}$$

And TF-IDF value was defined as:

$$w_{ij} = tf_{ij} \times \log(\frac{N}{df_i})$$

Where tf_{ij} is the frequency of the token i in the document j, df_i is the number of documents that contain the token i and N is the total number of documents per user.

For machine learning algorithms we used the implementations that are arranged in Python sklearn library and among them we have RandomForestClassifier, NearestCentroid and OneVsOneClassifier for the three classifiers used in the ensemble.

To determine the definitive class to which a set of documents belongs with the ensemble of classifiers, we use a majority voting method, which consist of considering as the class of the document that which has been predicted by the largest number of classifiers.

For the validation process we use the StratifiedKFold from sklearn.model_selection module to perform a 5-Strified-K-Fold validation whit the training corpus which is divided into train and test respectively to be able to evaluate the effectiveness of the system. As an evaluation metrics we use F1 score for Topic an Age dimensions and for Gender we use Accuracy score from sklearn library in the first run. For second run we use the two different rankings proposed in the task to evaluate the participants: ranking 1 which evaluate the performance of each system using a partial scoring scheme, giving 1/3 of the points for each correctly predicted profile and 0 points if neither is correct; and ranking 2 which gives 1 point only if all classes are well predicted and 0 otherwise.

3 Experiments and Results

The test dataset provided by the tasks organizers was similar to train corpus (which was unbalanced especially for gender class, with a predominance of male users), and it was composed by posts of 411 different users with unknown age, gender and topic classes.

To obtain the best possible results with our method, we realized several experiments varying the values of the parameters in order to determine a good configuration per class. At the end of the experimentation process, we choose two different runs to be presented. The first one (Team2_1_1), see in Table.1, has a different configuration per class according to the best obtained result in the individual classification. Age class has been represented with a configuration of 2-grams of characters, a 1000 feature reduction and with TF-IDF as the weight of features. Gender class has been represented with a configuration of 4-grams of characters, a 1000 feature reduction and TF as the weight of features and Topic class has been represented with 4-grams of characters, a 1000 feature reduction and TF-IDF as the weight of features.

Using the Strified-K-Fold Cross-Validation we obtain as a result of the individual evaluation per class 0.3732, 0.8854 and 0.7051 for age, gender and topic respectively.

In the second run (Team2_1_2), see in Table.1, we have adjusted the parameters to be the same in the three classes and use a single configuration in all: 4-grams of characters, a 1000 feature reduction and TF for the weight of features.

Using the two metrics given in the TAG-it page we evaluate the second run and obtain 0.6801 and 0.2914 for Metric 1 and Metric 2 respectively as result.

Run	Metric 1	Metric 2
Team1_1_3	0,6991	0,2506
Team1_2_3	0,6739	0,2433
Team1_3_3	0,6991	0,2506
Team2_1_1	0,4160	0,0924
Team2_1_2	0,4436	0,0924
Team3_1_1	0,6626	0,2530
Team3_1_2	0,7177	0,3090
Team3_1_3	0,7347	0,3309

Table.1 Competition results for subtask 1.

The results obtained were not as good as expected compared with the results obtained in the validation process that we made, considering that

n-gram of character representation obtained low scores for topic and age classification.

4 Conclusion

In this paper we described the proposal presented to participate in the TAG-it author profiling task from EVALITA 2020. Our proposal is based on an ensemble of machine learning algorithms with three well known classifiers and a Bag of Word of characters n-grams using a feature reduction by a predefined parameter and calculating TF or TF-IDF for features weight.

To resolve subtask 1 we proposed two different strategies where we first adjust the values of the parameters n for n-grams, k for feature reduction and Tf or TF-IDF for feature weight to the classification of each profile independently using a different configuration in each one, and in the second we just adjust a general parameter and use the same configuration in the three profiles classification at once.

Despite that the fact that in the evaluation process we carried out obtained better scores, the results of the task were not as good as expected, since low results were obtained for topic and gender dimension.

Reference

Francisco M. Rangel Pardo, Paolo Rosso: Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter. *CLEF (Working Notes)* 2019

Mario Ezra Aragón, Miguel Ángel Álvarez Carmona, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Daniela Moctezuma: Overview of MEX-A3T at IberLEF 2019: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. *IberLEF@SEPLN 2019*: 478-494

Miguel Á. Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, Antonio Rico-Sulayes: Overview of MEX-A3T at IberLEF 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. *IberLEF@SEPLN 2018*

Valdez-Rodríguez, J.E., Calvo, H., Felipe-Riverón, E.M.: Author profiling from images using 3d convolutional neural networks. In: *In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019)*, CERUR WS Proceedings (2019)

Daniel Castro Castro, Maria Fernanda Artigas Herold, Reynier Ortega Bueno, Rafael Muñoz: Cerpamid-UA at MexA3T 2019: Transition Point Proposal.

Cimino A., Dell´Oreleta F., Nissim M. (2020). "TAG-it@EVALITA2020: Overview of the Topic, Age, and Gender prediction task for italian". *In Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian.*