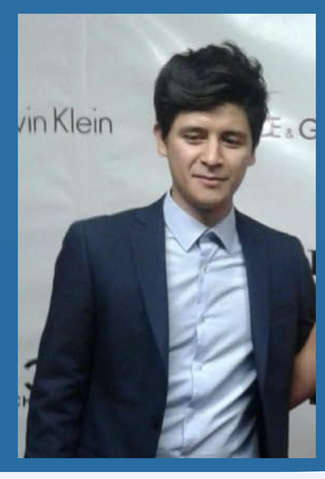
Bots and Gender Profiling using Character Bigrams

Daniel Yacob Espinosa, Helena Gómez-Adorno and Grigori Sidorov

Instituto Politécnico Nacional, Centro de Investigación en Computación
Universidad Nacional Autónoma de México,
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas



Introduction

Use supervised learning to make a clasification between bots and users and classify human users by the gender using only Twitter tweets.

The dataset to perform the classification is in English and English.

Each file is a user that contains 100 tweets with all the features that form a tweet (@Mensions, #Hashtags, Links and Emoticons).

Methods

Using a preprocessing of the data, we removed digits, Url, @Mentions, Emoticons, punctuation marks and finally characters that were not in the ASCII Standard.

After that preprocessing erase the spaces between words, that way we would have all the words together ready to use a structure.

For the task of finding the features for the classification we decided to use bi-grams of character.

When we have all bigramas structures of our texts it is necessary to organize them so it is necessary to use a vector space model.

using the frequency of bigramas by placing their respective numerical value within matrix for both problems.

Matrix	Doc1	Doc2	Doc3
bi-gram1	3	6	0
bi-gram2	0	3	2
bi-gram3	5	0	1

Table 1. Example of Term-Document Matrix with character bi-gram

We decided to use SVM (Support Vector Machine) as a classifier because of its class separation performance. We use 10-part cross Validation for training and evaluation, obtaining good results.

Results

We performed more tests for the formation of characteristics but with the bi- grams we obtained better results,

Classification method	Spanish			English		
	1-gram	2-grams	3-grams	1-gram	2-grams	3-grams
J48 Split 70%	62.00	67.82	82.03	63.44	65.54	71.29
NaiveBayes Split 70%	65.11	71.11	83.30	66.48	69.25	72.69
RandomForest Split 70%	91.77	74.44	85.11	83.21	85.55	83.24
RandomForest 20-Fold	92.22	92.7	90.6	92.79	92.80	90.13
SVM CrossValidation-10	91.86	**92.38**	90.76	92.42	**92.86**	90.41

Table 2. Evaluation results in terms ofthe classification accuracy between humans and bots on the PAN Author Profiling 2019 test corpus and classification method.

Classification method	Spanish			English		
	1-gram	2-grams	3-grams	1-gram	2-grams	3-grams
J48 Split 70%	53.22	57.1	56.31	54.44	57.02	55.90
NaiveBayes Split 70%	65.11	71.11	83.33	62.60	66.08	69.91
RandomForest Split 70%	71.77	74.44	85.111	66.08	78.86	75.65
RandomForest 20-Fold	76.35	75.53	77.42	73.22	76.82	76.31
SVM CrossValidation-10	76.13	**80.72**	72.40	75.22	**83.37**	78.39

Table 3. Evaluation results in terms ofthe classification accuracy of gender on the PAN AuthorProfiling 2019 test corpus and classification method

Conclusions

Using the main feature bigrams for the classification between users and bots was very good, additionally we can say that you should use features of the social network or environment where you want to classify since they can serve much better for the formation of classes within the classifiers.

Therefore, for gender classification, there are no bad results with the formation of bigrams but it would be interesting to try other types of structures for classification.