Author Profiling Using Support Vector Machines Notebook for PAN at CLEF 2016

Rodwan Bakkar Deyab, José Duarte, Teresa Gonçalves

d34642@alunos.uevora.pt, d10401@alunos.uevora.pt, tcg@uevora.pt

Introduction

Author profiling problem is about detecting some characteristics (age, gender, for example) of the author of some text depending on the features (eg. lexical, syntactical) of this text. Men and women, and of different ages, write in different ways. Having a dataset in hand, written by different authors of different characteristics, we can train the machine using this dataset so it can predict these characteristics of an unseen piece of text fed to it. PAN16 author profiling task provides a dataset of tweets for the sake of developing an author profiling system. The task is about predicting the age and the gender of the author. Machine Learning technique suits to achieve this goal. Support Vector Machines [1] (SVMs) can be used as a multi-class classifier which could be trained using the dataset provided to produce a model which can be consulted on an unseen set of tweets written by some author to predict his age and gender. Bag-of-Words [2] (BOW) is a simplified representation of the text corpus which contains all the words used in it with their frequencies. BOW representation is used in many areas like Natural Language Processing Information Retrieval, Document Classification and among others. In our work we use SVMs and BOW representation. We use the python machine learning library, scikit-learn [3]. After we produced the best possible model trained on PAN16 author profiling dataset, we ran some tests over the test sets provided by Tira.

Dataset

The dataset provided by the PAN 2016 was used in our study. The corpus contains 436 files, each file contains a set of tweets and these files are written by different authors. Table 1 explains this dataset.

Gender	Age Range	Number of files	Total
Females	18-24	14 (3%)	218
	25-34	70 (16%)	
	35-49	91 (20%)	
	50-64	40 (9%)	
	65-xx	3 (0.6%)	
Males	18-24	14 (3%)	218
	25-34	70 (16%)	
	35-49	91 (20%)	
	50-64	40 (9%)	
	65-xx	3 (0.6%)	
	Total		436

Table 1: PAN 16 Author Profiling Dataset

Results

We show one of the results after testing on the PAN 16 corpus with 10000 features. Table 2 shows the results:

	precision	recall	f1-score	support	kernel	gamma	c
class1	0.00	0.00	0.00	2	rbf	0.0001	100
class2	0.22	0.15	0.18	26			
class3	0.31	0.56	0.39	27			
class4	0.25	0.08	0.12	12			
class5	0.00	0.00	0.00	1			
class6	0.00	0.00	0.00	2			
class7	0.43	0.46	0.44	26			
class8	0.29	0.29	0.29	34			
class9	0.33	0.21	0.26	14			
avg / total	0.30	0.31	0.29	144			

Table 2: Results for PAN 16 corpus with 10000 features.

Conclusion and Future Work

We notice that the results were better for PAN15 than for PAN16 and that was because of the tagging process worked differently for different dataasets. Increasing the number of features does not mean necessarily better results. We can improve our results by adding features extracted with respect to the natural language (syntactic and semantic features, for example).
In the future, the use of the term frequency-inverse document frequency (tf-idf) technique and tuning the maximum size of the BOW can help to improve the results.

Acknowledgements

We want to address our thanks to the Departamento de Informática da Escola de Ciências e Tecnologia da Universidade de Évora, for all the support to our work.

System Architecture

Our system has three modules: preprocessing, training and testing modules. In figure 1 we show the architecture of the system in the training phase. In figure 2 we present the architecture of the system in the testing phase. Both of them use the preprocessing module.

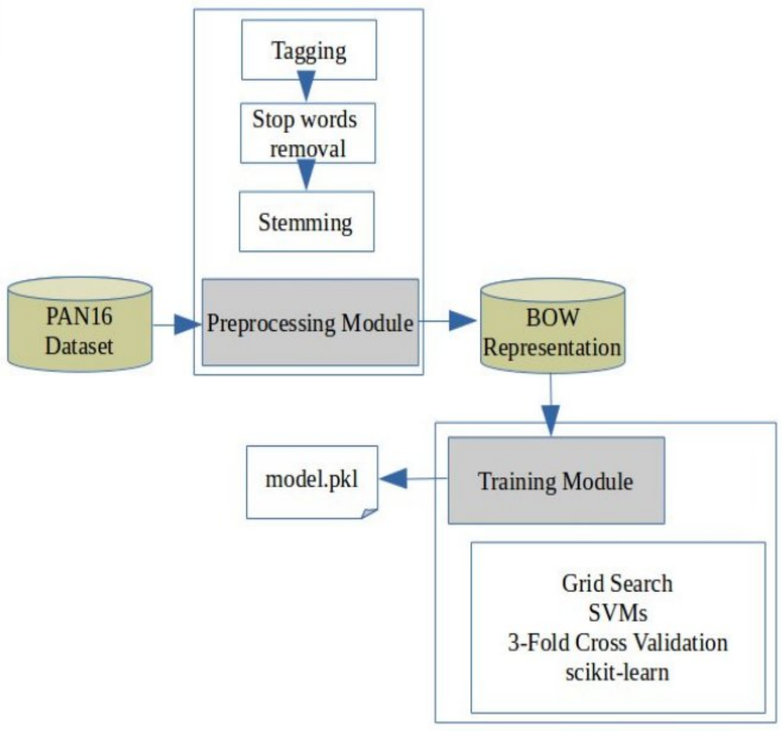

Figure 1: The architecture of the system: training phase

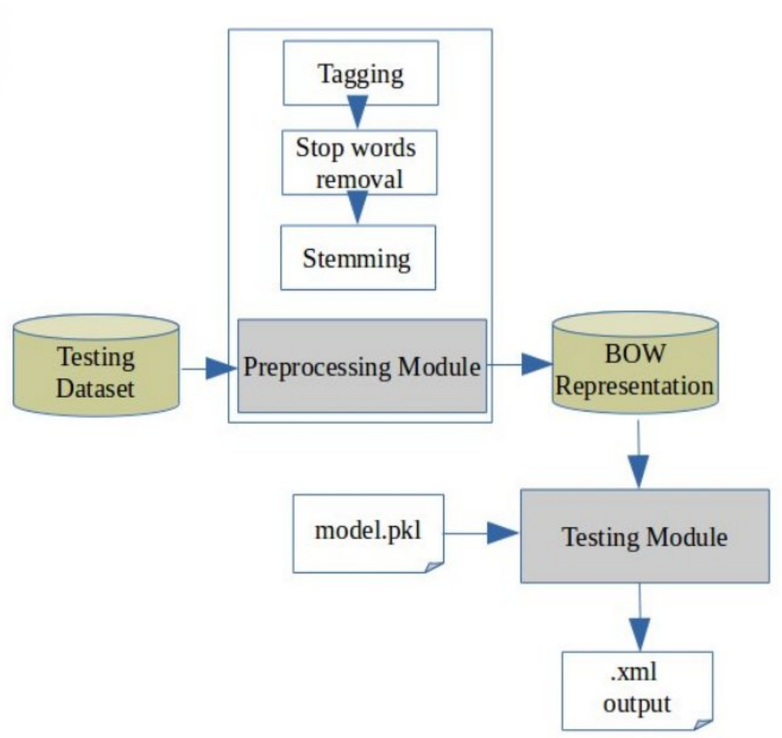

Figure 2: The architecture of the system: testing phase

References

[1]: Hearst, Marti A. and Dumais, Susan T and Osman, Edgar and Platt, John and Scholkopf, Bernhard: Support vector machines. Intelligent Systems and their Applications, IEEE 13(4), 18–28(1998)

[2]: Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics 1(1-4), 43–52 (2010)

[3]: 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, 2825–2830 (2011)