Learning from data final project

Chris Pool

2016/01/19

Contents

1 Introduction

The final assignment for the course Learning from data was to build a model to predict gender and age category given multiple tweets from several users. The task proved to be challenging because of the limited amount of time and data. In this report you can find how I have build a system to classify the tweets and predict the user's age and gender. The evaluation can be found in the last chapter. While writing this report my system will be tested on a test set, those results will be presented during the final presentations.

2 Data

The data consisted of several tweets of multiple users in four languages. The data was fairly even divided by age, in general there where a few more tweets from male users. For age classification the data was a bit more skewed as can be seen in table 2.

Table 1: Tweets divided by gender

	Male tweets	Female tweets	Total tweets
English	5013	4935	9948
Dutch	1000	776	1776
Italian	1400	1187	2587
Spanish	3647	3291	6938

Table 2: Tweets per age category

	18-24 tweets	25-34 tweets	35-49 tweets	50-XX tweets	Total tweets
English	4175	3554	1392	827	9948
Spanish	2868	2000	1470	600	6938

For Dutch and Italian there is no data for age classification. Also the amount of tweets for proper gender classification is a bit low.

3 Approach

The first step was to build a framework where I could experiment with different classifiers and feature extraction methods. I used Sci-kit learn (Pedregosa et al. 2011) for this combined with some functions from NLTK (Bird, Klein, and Loper 2009).

My system can be called with two arguments, the first one, the location of the training set, is required. The second, the location of the test set, is optional. If a test set is not supplied my system goes into test mode, meaning that the training data is split in a train set (75%) and a development set(25 %) that I can use for evaluation. If the second argument is supplied means that the system is not running in test mode and it is not possible to evaluate the results because the gold labels are unknown and therefore writes the results to the truth files to be evaluated by in this case the instructors of this course.

The first step in the framework was to read the data and to pre-process the data as can be read in the next subsection. With the cleaned tweets I extracted the features from the data.

I have chosen to build a model for each sub-task, so for gender classification for each language and age classification for English and Spanish resulting in six separate models.

With the predictions of the models I calculate the labels for the user based on the most common label in the prediction. E.g. if 20 of the 30 tweets of the user are classified as Female I assign the Female label to that user.

Depending on the mode the system is running in a evaluation is done based on the gold standard or the results are written to the truth.txt files.

3.1 Pre-processing

The tweets contain a lot of noise, some of the noise could be useful information for the classifiers so I did not change to much:

- Convert links to URL
- Convert username to AT_USER
- Convert each number to NUMBER
- Remove additional white spaces
- Standardize floating characters to three characters (e.g. hellooooo to hellooo)

Some noise that could be relevant are capital letters in the text and other punctuation so I did not remove them.

3.2 Classifiers

For all tasks I used a Support Vector Machine (SVM) classifier. This classifier performed the best compared to Naive Bayes and a decision tree. I tried several settings for the classifier

The results are further explained in the evaluation chapter but in the table below the average results per classifier can be found for the gender classification task. The classifier in bold is the classifier I used.

Table 3: Classifiers used for gender classification

Classifier	Precision	Recall	F-score	Accuracy
SVM Linear kernel C = 1.0	**0.81**	**0.81**	**0.81**	**0.81**
SVM linear kernel C = 0.5	0.80	0.81	0.81	0.81
Naive Bayes	0.72	0.72	0.72	0.72
Decision tree	0.77	0.77	0.77	0.77

For age classification the results can be found in table 4.

Table 4: Classifiers used for age classification

Classifier	Precision	Recall	F-score	Accuracy
SVC (kernel='rbf', C=1.0, gamma=0.9)	**0.46**	**0.58**	**0.49**	**0.58**
SVC (kernel='rbf', C=1.0, gamma=0.5)	0.48	0.56	0.48	0.56
Naive Bayes	0.48	0.56	0.47	0.56
Decision tree	0.58	0.53	0.46	0.53

Table 5: Features used and corresponding weights

	English Gender	Dutch Gender	Italian Gender	Spanish Gender	English Age	Spanish Age
Repeating letters	1				1	1
Capped words					1	1
Capital letters						1
Language specific	1					
Latin			1	1		
Mentions	1					
Hashtags	1					
Bag of words	2	2	2		1	1
Character n-grams	2	2	2	2		1

3.3 Features

I used several features for the different classifiers. Most of these features are based on looking at the data. Character n-grams and repeating letters are features used in the entry of the pan 2015 shared task by (Grivas, Krithara, and Giannakopoulos n.d.) and had a positive effect. Sci-kit learn has the option to assign weights to the different feature extraction methods. In table 5 these weights can be found.

Features used:

Repeating letters If repeating characters (3 or more) are used, e.g. hellooooo becomes hellooo

Capped words Number of words that start with capital letter.

Capital letters Total number of capital letters

Language specific List of words for each language that possibly indicates gender/age

Latin Amount of words ending with 'a' that indicates female word in Spanish and Italian.

Mentions Number of mentions

Hashtags Number of hashtags

Bag of words Tf-IDF using all words

Character of n-grams Tf-IDF of character 3-grams.

I also experimented with other features like using the gender classifier as feature for the age classifier but that did not improve classification. Also counting punctuation did not improve the result.

4 Evaluation

For each task the precision, recall, f-score and accuracy is calculated. The task is to predict for each user the gender and age but for completeness I also mention the evaluation per tweet. The evaluation below is based on running my system on the train set that is split in a train and test set (75%/25%)

4.1 Gender Classification

The gender classification is done on a tweet level, for each user all tweets are classified, the label assigned to the user is the one that is most frequent in the prediction. There is a large difference between the accuracy on a tweet level and user level. I think this has to do with the fact that only a few tweets for each user has information about the gender and a lot of tweets are gender neutral.

Table 6: Results per tweet

	Precision	Recall	F-score	Accuracy	Support (n_tweets)
English	0.59	0.58	0.58	0.58	2487
Dutch	0.60	0.60	0.60	0.60	444
Italian	0.62	0.61	0.61	0.61	647
Spanish	0.61	0.60	0.60	0.60	1735
Total/avg	**0.60**	**0.59**	**0.60**	**0.60**	**5313**

Table 7: Results per user

	Precision	Recall	F-score	Accuracy	Support (n_users)
English	0.81	0.78	0.78	0.78	27
Dutch	0.85	0.80	0.78	0.80	5
Italian	0.71	0.71	0.71	0.71	7
Spanish	0.89	0.89	0.89	0.89	18
Total/avg	**0.81**	**0.81**	**0.81**	**0.81**	**57**

Compared to the systems that participated in the PAN 2015 shared task the results seem pretty good, I expect that the results on the test set are a bit lower because there where some duplicate tweets in the training set causing over fitting.

4.2 Age Classification

The age classification proved to be more difficult. Especially the Spanish model performed poorly. If I regard the baseline being the most frequent label then the baseline is an accuracy of 37%

Table 8: Results per tweet

	Precision	Recall	F-score	Accuracy	Support (n_tweets)
English	0.63	0.58	0.55	0.58	2487
Spanish	0.46	0.37	0.33	0.37	444
Total/avg	**0.00**	**0.00**	**0.00**	**0.00**	**5313**

Table 9: Results per user

	Precision	Recall	F-score	Accuracy	Support (n_users)
English	0.62	0.67	0.61	0.67	27
Spanish	0.21	0.33	0.24	0.34	18
Total/avg	**0.44**	**0.53**	**0.46**	**0.53**	**45**

4.3 Overall results

To compare my system with other systems I calculated the average accuracy for all tasks and languages.

Table 10: Final results

Team	Global	English	Dutch	Italian	Spanish
alvarezcarmona15	0.84	0.79	0.82	0.81	0.91
My system	**0.67**	**0.73**	**0.80 ***	**0.71***	**0.6**

* only gender accuracy

The winner of last year, the team *AlvarezCarmona15*, performed much better than my system. The biggest difference was the Spanish age task. To compare the systems is not completely fair because our system only did age classification for Spanish and English.

5 Discussion

To improve my system more time is needed to evaluate exactly which features contribute the most. I did not find a method to test all possible features and let Sci-kit learn detect the best combination of features.

Also some features need to be improved. In Spanish and Italian some words ending with an "a" indicate a female sense, so doing POS-tagging combined with some rules or a list based approach would probably be a useful feature.

The language specific feature could also be improved, the time I have until my presentation I want to try to improve this feature.

Another improvement could be determining the label of the user. I choose for a method of the most frequent label in the prediction but another option is counting the probability for each label. I saw some cases where the probability was very low and maybe it is better to ignore those tweets.

References

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media.

Grivas, Andreas, Anastasia Krithara, and George Giannakopoulos. "Author profiling using stylometric and structural feature groupings". In:

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.