A supervised method for cleanup of a User-Generated Content corpus

Enrique Manjavacas[1], Ben Verhoeven[2], and Walter Daelemans[2]

[1]Free University of Berlin

[2]CLiPS - University of Antwerp

Summary

• In this project, we evaluate machine learning methods for cleaning up noisy data using a self-compiled corpus of weblog text which is meant to be used in stylometry research.

The use of blog texts, as is common in current stylometry research, usually leads to the incorporation of observations that are noisy with respect to the object of study (e.g. *gender*). This poses the theoretical problem of to what extent systems are learning real style differences. We investigate the possibility of performing an automatic clean-up of the data.

Dataset Description

We have made use of a self-compiled blog corpus that was created from the domains `http://www.blogger.com` and `http://www.blogspot.com` in the months of October and November 2014. Statistics about the corpus are shown in Figures 1 and 2.

Words	BE	NL	DE	AT
Gender	23,126,338	57,532,099	74,944,965	21,174,985
Profession	8,198,389	2,522,961	2,665,512	255,939
Location	22,282,351	6,520,240	8,839,919	2,216,072

Fig. 1: Number of words per country

Blogs	BE	NL	DE	AT
Gender	1571	3183	4235	1726
Profession	548	124	105	39
Location	1196	399	412	260

Fig. 2: Number of blogs per country

• We randomly sampled 550 blogs from the NL section of our blog corpus.

• Only one annotator was responsible for labeling and instructions were provided to annotate according to the following definition of the negative class:

Blogpost redirection as evidenced in "community blogs".

Non-authored content as in the case of "filter blogs" - a concept by Susan C. Herring.

Unchecked blog profiles, as often encountered in "corporate blogs".

• Profiles with both known and unknown gender were included in the sample to investigate the usefulness of gender information for cleaning purposes.

As shown in the table, a strategy of excluding blogs with unknown gender could in principle only reduce the proportion of noisy data from 20.88% to 15.73%.
The resulting class distribution, depicted in Figure 3, shows a strong association of the negative class (irrelevant blogs) with male bloggers.

	Relevant	Irrelevant	Total
Female	310	36	346
Male	87	40	127
Unknown	21	37	58
Total	428	113	541

Fig. 3: Class proportions in the annotated dataset

Classification Setup

Feature Construction We extracted token bigrams and trigrams and character trigrams and tetragrams. As a feature selection method, we selected the 1000 and 5000 most frequent items for each feature set.

Feature Representation Feature extraction was conducted according to both the "relative frequency" and "binary representation" of the features in each document.

Classifier Support Vector Machines Gaussian kernel as implemented in the library `libSVM` were used. This decision was motivated by its good off-the-shelf performance which also suits the exploratory character of this classification task. Linear kernels were discarded due to general inferior performance on our data.

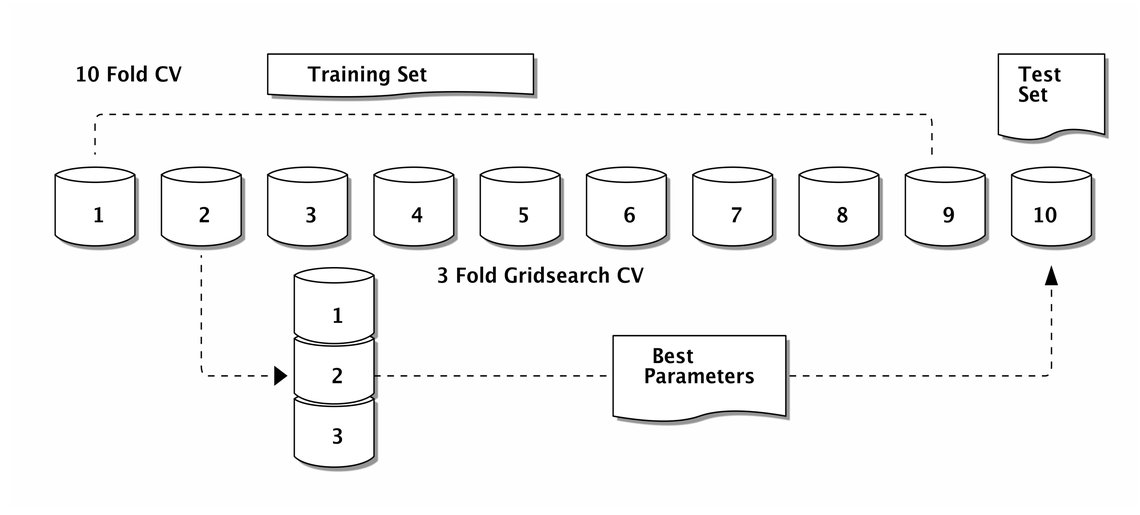

Fig. 4: Iteration in a Nested cross-validation setup

Nested cross-validation We used a nested cross-validation loop as depicted in Figure 4.

Grid Search was used in the inner loop with a main focus on achieving high *precision* of the positive class. This is motivated by the practical goal of the corpus filtering where false negatives are less harmful than false positives.

Evaluation

Cross-validation provided a distribution of precision, recall and F-measure scores for each of the feature sets. Average score and standard deviation were computed and compared to the baseline classifier.

Baseline was set to be the majority class in the entire sample which in our dataset was represented by the positive class (relevant blog data) with a proportion of 79.23% (See Figure 3).

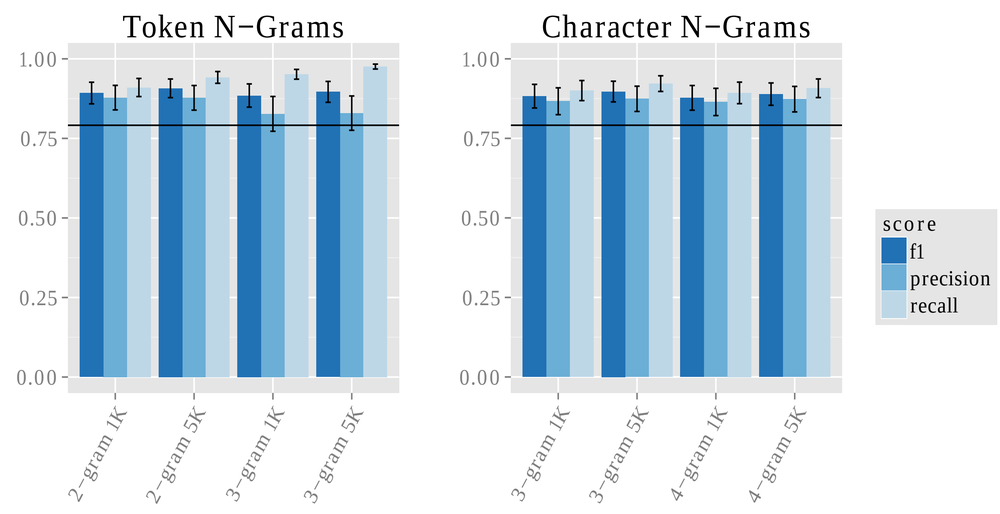

Fig. 5: Evaluation of the classifiers

Figure 5 summarizes the classification scores for the different feature combinations. The baseline is plotted as a horizontal line. From the results we can draw the following conclusions:

Relative-frequency performed better than binary representation, which was for this reason soon dropped from the model.

Character n-grams performed best and showed robustness in score across both number of features and n-gram order.

Token n-grams displayed larger variance than character n-grams and their accuracy dropped eventually under the baseline with increasing n-gram order.

Number of features did not increase accuracy in general but it seemed to be associated with higher recall. This was to be interpreted carefully since higher recall of the positive class, accompanied by a drop in the scores of the negative class, indicated a tendency towards overfitting.

Work in progress

Although the classifier is performing better than the baseline, there is still room for substantial improvements. More concretely, follow-up research will address the following issues:

• A fine-grained subclassing of the negative class (irrelevant blogs). The current negative class has an inherently heterogenous character and might be better learned in a multiclass setup.

• The use of more feature selection techniques from IR instead of frequency based selection. Specifically *tf-idf* and χ^2 feature selection might result in a better fit of features to the classes.

• The efficiency of the classifier needs to be tested against raw new data and its output manually and qualitatively evaluated against our definition of noisy data. We hypothesize that stylometry research can profit from such a pre-training filtering phase. Follow-up research project will attempt to track eventual significant performance differences caused by the aforementioned filtering phase. We hypothesize that the data resulting from filtering will show higher homogeneity of classes and will also constitute a better sample of real differences in gender style.

Follow-up Research

Previous research in gender profiling by Argamon & Koppel has used datasets similar to the one presented here. Although changes in the dynamics of the blogosphere have taken place, one might expect to encounter a similar proportion of noisy data in their blog corpus.

Style differences Given the association between noisy blogs and male blog ownership, one could be tempted to hypothesize that the learned classifiers were picking up on the blog types associated with our negative class instead of real gender differences in language style.