NLP-based Feature Extraction for Automated Tweet Classification

Anna Stavrianou, Caroline Brun, Tomi Silander, Claude Roux

Xerox Research Centre Europe, Meylan, France

Name.surname@xrce.xerox.com

1 Introduction

Traditional NLP techniques cannot alone deal with twitter text that often does not follow basic syntactic rules. We show that hybrid methods could result in a more efficient analysis of twitter posts. Tweets regarding politicians have been annotated with two categories: the opinion polarity and the topic (10 predefined topics). Our contributions are on automated tweet classification of political tweets.

2 Combination of NLP and Machine Learning Techniques

Initially we used our syntactic parser [1] which has given high results on opinion mining when applied to product reviews [2] or the Semeval 2014 Sentiment Analysis Task [3]. However, when applied to Twitter posts, results were not satisfactory. Thus, we use a hybrid method and combine knowledge given by our parser with learning.

Linguistic information has been extracted from every annotated tweet. We have used features such as bag of words, bigrams, decomposed hashtags, negation, opinions, etc. The"liblinear" library (http://www.csie.ntu.edu.tw/~cjlin/liblinear/) was used to classify tweets. We used logistic regression classifier (with L2-regularization), where each class c has a separate vector w_c of weights for all the input features. More formally, $P(c|x; w_c) \propto e^{\sum_{i=1}^{d} w_{ci} x_i}$, where x_i is the ith feature and the w_{ci} is its weight in class c. When learning the model, we try to find the vectors of weight w_c that maximize the product of the class probabilities in the training data.

Our objective has been to identify the optimal combination of features that yields good prediction results, while avoiding overfitting. Some features used are: Snippets: during annotation, we kept track of the snippets that explained why the annotator tagged the post with a specific topic or polarity, Hashtags: decomposition techniques have been applied to hashtags, and they are analyzed by an opinion detection system that extracts the semantic information they carry [4].

We have selected the models using a 10-fold cross validation in the training data and evaluated them by their accuracy in the test data. For the topic-category task, (6,142 tweets, 80% used for training), the annotation had <0.4 inter-annotator agreement, which shows the difficulty of the task. Table 1. shows the results when NLP features are used, as well as when some semantic merging of classes takes place.

In: P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, Y. Toussaint (Eds.): Proceedings of DMNLP, Workshop at ECML/PKDD, Nancy, France, 2014.

Table 1. Cross-validation (2nd col) and prediction (3rd col) results for topic classification.

NLP features	44.38	29.37
NLP features + merging	48.91	34.17

Binary classification was applied to improve the results. We selected the class with the highest distribution and annotated the dataset with CLASS1 and NOT_CLASS1 tags. We created a model for the prediction of CLASS1, the prediction of CLASS2 and a model for the prediction of the rest of the 8 classes. Merging these models gave an accuracy of 40.03%, higher than the max accuracy of Table 1.

Table 2. Binary classification results (2nd col: cross-validation, 3rd col: prediction) for topic.

CLASS1/NOT_CLASS1	85.28	62.57
CLASS2/NOT_CLASS2 (removal of CLASS1)	92.10	68.42
The rest of the classes (removal of CLASS2)	49.58	38.24

For the opinion polarity task (5,754 tweets, 80% used for training), the inter-annotator agreement was higher (~ 0.8). As Table 3. shows, we have used not only NLP features from the tweet but also from the 'snippet'. The "syntactic analysis" is the opinion tag given from our opinion analyser.

Table 3. Cross-validation (2nd col) and prediction (3rd col) results for the opinion polarities.

NLP features (syntactic analysis of opinion)	61.28 (62.13)	56.77 (56.6)
NLP features of snippet (syntactic analysis)	66.41(67.99)	61.2 (61.46)

As a conclusion, in this paper we provide a model that predicts opinions and topics for a tweet in the political context. More research around feature analysis will be carried out. We also plan to add more features yielded by our syntactic analyzer such as POS tags, or tense. We should also consider a multiple-class labelling.

3 Acknowledgements

This work was partially funded by the project ImagiWeb ANR-2012-CORD-002-01.

4 References

1. Ait-Mokthar, S., Chanod, J.P.: Robustness beyond Shallowness: Incremental Dependency Parsing. NLE Journal, 2002.
2. Brun, C.: Learning opinionated patterns for contextual opinion detection. COLING 2012.
3. Brun, C., Popa, D., Roux, C.: XRCE: Hybrid Classification for Aspect-based Sentiment Analysis. In International Workshop on Semantic Evaluation (SemEval), 2014 (to appear).
4. Brun, C., Roux, C.: Décomposition des « hash tags » pour l'amélioration de la classification en polarité des « tweets ». In TALN, July, 2014.