Stylometric Features for Emotion Level Classification in News Related Blogs

Elisabeth Lex
Know-Center GmbH
Inffeldgasse 21a
8010 Graz, Austria
elex@know-center.at

Michael Granitzer
Know-Center GmbH
Inffeldgasse 21a
8010 Graz, Austria
mgrani@know-center.at

Markus Muhr
Know-Center GmbH
Inffeldgasse 21a
8010 Graz, Austria
mmuhr@know-center.at

Andreas Juffinger
The European Library
c/o the Koninklijke Bibliotheek
2509 LK The Hague
andreas.juffinger@kb.nl

ABSTRACT

Breaking news and events are often posted in the blogosphere before they are published by any media agency. Therefore, the blogosphere is a valuable resource for news-related blog analysis. However, it is crucial to first sort out news-unrelated content like personal diaries or advertising blogs. Besides, there are different levels of emotionality or involvement which bias the news information to a certain extent. In our work, we evaluate topic-independent stylometric features to classify blogs into news versus rest and to assess the emotionality in these blogs. We apply several text classifiers to determine the best performing combination of features and algorithms. Our experiments revealed that with simple style features, blogs can be classified into news versus rest and their emotionality can be assessed with accuracy values of almost 80%.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Web Content, Blogosphere, Data Mining; I.5.2 [**Feature Evaluation and detection**]: Feature evaluation and selection

General Terms

Algorithms, Experimentation, Performance

Keywords

Blogosphere, Classification, News

1. INTRODUCTION

In 2008, the popular blog search engine Technorati analyzed its index data regarding the most popular tags and reported that in June, *News* was the number one most-used tag by bloggers, appearing close to 200,000 times over the course of the month[1]. This emphasizes that a lot of blog content is related to news and events. Therefore, news related blogs are an excellent resource to analyze trends, opinions and different aspects of news stories over time. However, the automatic identification of news related blogs is still challenging. In our work, we address this challenge with a genre detection [1] related approach to classify blogs into news related versus rest. We refer to this as news versus rest (NvR) task. We define news related blogs as written by an individual or a group of people independent of any news agency, with the primary intention of commenting current events prepared for some community of relevant size. Furthermore, we assess the emotionality within the news related blogs in order to identify the feelings of individuals toward specific events. Especially for media resonance analysis, it is highly important to determine the reaction towards certain events or campaigns. If an author blogs emotionally, the event definitely concerns her and therefore this event naturally attracts more attention. This is further referred to as emotion assessment (EA) task.

2. DATASET AND FEATURES

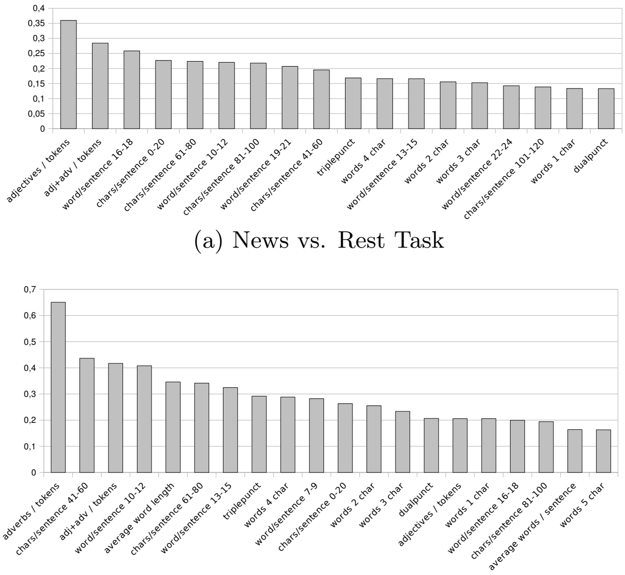
Our dataset consists of a randomly selected subset of the TREC Blogs08 Dataset[2]. We manually annotated 83 blogs with total number of 12844 distinct blog entries in English into *news* versus *non-news* and into *emotional*, versus *not emotional.* Note that all our experiments are done on blog post level. As features, we exploited stylometric features [2] because they are inherently topic independent since style does not depend on topics. Such features provide us with a high degree of generalizability in the inhomogenous topic landscape of the blogosphere while being simple. As features, we considered the punctuation distribution, emoticons, words per sentence, characters per sentence, noun-verb groups per sentence, the average number of unique POS tags per sentence, the ratio of lower to upper case characters, the word length distribution, the adjective rate, and the adverb

RIAO'2010, April 28-30th, 2010, Paris, France

[1] http://technorati.com/blogging/state-of-the-blogosphere/the-how-of-blogging/

[2] http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html

rate. To identify the most relevant features, we calculated the linear correlation (LC) of the features with the categories news and emotionality. Figure 1(a) and 1(b) show the LC of the top 20 stylometric features. For the NvR task (Figure 1(a)), the LC reveals that the *adjectives/token* feature is the most correlated with about 0.36. The feature *adjectives + adverbs / tokens* also has a very high correlation due to the linear dependency on the first feature. Besides, the sentence length, and the sentence complexity (number of words per sentence) are highly correlated features for the NvR task.



(a) News vs. Rest Task

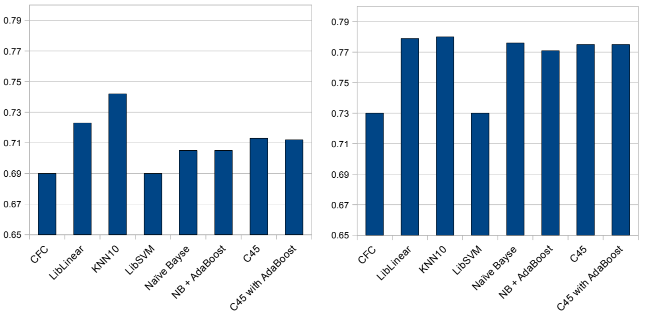
(b) Emotion Assessment Task

Figure 1: Linear Feature Correlation

For the EA task, the *adverbs/token* feature is most correlated. Different from the NvR task, the number of used adverbs is much more relevant for this task. An interpretation might be that an individual who writes very emotionally often uses many adverbs whereas a rational writer uses adverbs rarely. The eight best feature is the triple punctuation feature which covers emoticons often used in blogs.

3. CLASSIFICATION

As text classifiers, we used a Support Vector Machine (SVM)) based on LibLinear as well as an SVM based on LibSVM, a k-NN algorithm with $k = 10$, and a centroid-based text classifier, the Class Feature Centroid (CFC) with $b = 1.1$ [3]. Also, a Naive Bayes classifier, with and without boosting (AdaBoost) as well as a C45 decision tree with and without boosting (AdaBoost), all taken from Mallet[3]. Experiments with 10-fold crossvalidation revealed that the best classification results in an accuracy of 78% (C45 Boosted) for the EA task and an accuracy of 75% (k-NN) for the NvR task. Note that for these experiments, we took the features with highest linear correlation into account. We used the adjective ratio, the word number per sentence distribution, the word length distribution and the emoticon count as input for the classifiers. The results for the different classifiers are shown in Figure 2. Even though we conducted a parameter search for the LibSVM which resulted in a performance increase of 2%, the LibSVM performs worse than other algorithms. One reason for this is that we did not perform any feature normalization tasks although the features are of different scale. This also explains why the k-NN performs best in both tasks. As our experiments show, the classifier performance is limited with stylometric features, especially when compared to typical high dimensional features spaces based on nouns, unigrams, or stems. For instance, on stems, we achieved an accuracy of around 90% for both tasks. Nevertheless, stylometric features are guaranteed to be topic independent and therefore their generalization capabilities should be much better in the blogosphere.



(a) News vs Rest Task

(b) Emotion Assessment Task

Figure 2: Stylometric Features: Classification Accuracy

4. CONCLUSIONS

In this work, we investigated stylometric features to determine the best performing features and classifiers for the news versus rest (NvR) and the emotionality assignment (EA) task. The aim of the NvR task is to assign blogs to the news genre whereas in the EA task, the emotionality in these blogs is assessed. We computed the linear correlation to identify the most relevant features for our tasks. Our experiments revealed that with simple topic independent stylometric features, an accuracy of 78% can be achieved for the EA task and 75% for the NvR task.

5. ACKNOWLEDGMENTS

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

6. REFERENCES

[1] G. C. K. Chul Su Lim, Kong Joo Lee. Automatic genre detection of web documents. *Lecture Notes in Computer Science*, 3248:310–319, 2005.

[2] J. Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 2007.

[3] H. Guan, J. Zhou, and M. Guo. A class-feature-centroid classifier for text categorization. In *Proc. Int. Conf. on World Wide Web (WWW)*, New York, NY, USA, 2009.

[3] `http://mallet.cs.umass.edu/`