A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs

Elisabeth Lex
Know-Center GmbH
Graz, Austria
elex@know-center.at

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

Michael Granitzer
Know-Center GmbH
University of Technology Graz
Graz, Austria
mgrani@know-center.at

***Abstract*—In the blogosphere, the amount of digital content is expanding and for search engines, new challenges have been imposed. Due to the changing information need, automatic methods are needed to support blog search users to filter information by different facets. In our work, we aim to support blog search with genre and facet information. Since we focus on the news genre, our approach is to classify blogs into news versus rest. Also, we assess the emotionality facet in news related blogs to enable users to identify people's feelings towards specific events. Our approach is to evaluate the performance of text classifiers with lexical and stylometric features to determine the best performing combination for our tasks. Our experiments on a subset of the TREC Blogs08 dataset reveal that classifiers trained on lexical features perform consistently better than classifiers trained on the best stylometric features.**

***Keywords*-Document Classification; Data Mining; Features**

I. Introduction

On the Web, a huge amount of information is available and navigating through it is not an easy task. Especially since the advent of Web 2.0, the amount of content is expanding and for search engines, new challenges have arisen to cope with the changing information need. As mentioned in [1], the problem of answering questions is transformed from *"finding a needle in a haystack to a process of being presented with a variety of needles and choosing the one you want"*. Therefore, providing different facets for the users' different information needs is crucial.

In this work, we aim to assign blogs to the news genre and to assess the emotionality in these blogs. Our goal is to support blog retrieval with genre information and facets. Users should have the possibility to filter news related blogs by emotionality since in some cases, they may be interested in e.g. emotional eye witness accounts to current events like the Haiti earthquake and in other cases, they may want to read neutral blogs about political events.

Our approach is to investigate which features and classifiers can be exploited to assign blogs to the news genre and to classify the news related blogs into emotional versus neutral. We compare standard lexical text classification features with stylometric features to identify a setting with high generalization ability.

Our contribution is therefore:

- We apply lexical versus stylometric/shallow text feature to separate news related blogs from the rest
- We classify the news related blogs into emotional versus neutral
- We propose a feature study on 10 lexical features with three state of the art classifiers
- We analyze the statistical properties of 10 stylometric features and perform a matrix evaluation with eight classifiers

II. Related Work

The automatic identification of news related blogs is still challenging. More specifically, the investigation of significant features to assess reliable news related content is important. The identification of news related blogs is related to genre detection [2]. For web genre detection, various features can be exploited. Note that genre detection should not be confused with the issue of identifying the subject or topic of documents [3] because most subjects occur in a variety of genres, even tough some topics may occur only in certain genres. Lim et al. describe in [4] five distinct sets of features to automatically classify the genre of web documents. They exploit the URL, HTML tags, and token information, lexical information and structural information. However, usually little genre information is encoded in the URL of blogs. Within blogs, usually links to other Web sites or other blogs are available. Yet this link structure is a measure for popularity and quality and there is no evidence that the link structure differs between news related blogs and other blogs. Also, HTML tags and page impress features, such as the amount of advertising per page, cannot be exploited because there is no evidence that e.g. the amount of advertisement is different over various genres.

Emotion classification is related to the field of sentiment classification where the goal is to classify text into positive versus negative. In [5], online news articles are classified into reader-emotion categories exploiting an emotion lexicon. Chesley et al. [6] extract textual and linguistic features and use a classifier to categorize blog posts with respect

to sentiment. This paper is an extension to [7] where we consider emotion classification as a two class classification problem, neutral versus emotional.

III. Approach and Features

Our approach is to classify blogs according to their genre into two categories, news related blogs versus rest. In the retrieved news related blogs, we assess the emotionality also with a supervised classification strategy. We applied the following text classification algorithms: a Support Vector Machine (SVM) based on LibLinear [8], a k-NN algorithm introduced by Aha et al. [9] with $k = 10$ and cosine similarity and a centroid-based text classifier, the Class Feature Centroid (CFC) with $b = 1.1$ which was recently introduced by Guan et al. [10]. We selected those algorithms, because on the one hand, SVMs and k-NN are standard text classifiers and on the other hand, the CFC provides a term weighting scheme that implicitly favors the most discriminant terms. Also, the CFC is known to outperform SVM for certain datasets.

Naturally, a number of different features and classification algorithms can be used for text classification. In our approach, we used topic independent stylometric features as well as common lexical features [11] since both are easy to extract and simple.

A. Stylometric features

The use of stylometric features grounds on the theory of style by Sanders[12]: *"Style is the result of choices made by an author within a context from a range of possibilities offered by the language system'*. Stylometry expects that stylistic variations depend on: Firstly, the author preferences and competence, secondly the genre, thirdly the communicative context, and lastly, the expected characteristics of the intended audience. The main application of stylometric analysis is by far the authorship identification [13].

For both our classification tasks, we selected a number of stylometric features with focus on topic independence since text classifiers can easily overfit to topics due to the natural correlation between topics and genres [3]. Since topics change rapidly on the Web, our goal is to create a classifier independent of topics. Therefore, we exploit the following topic independent stylometric features which we first introduced in [7]:

- *Punctuation* The punctuation distribution, the count of one of 12 punctuations per document
- *Emoticons* The average number of sequential 2,3, and n punctuations
- *Words in sentences* The distribution of sentences with word length 0-3,4-6,7-9,10-12,13-15,...
- *Avg words / sentences* The average number of words per sentence
- *Chars in sentences* The distribution of sentences with 0-20,21-40,41-60,61-80,81-100,... characters
- *Avg chars / sentences* The average number of characters per sentence
- *Noun+verb sentences* The average number of minimal (in)correct sentences
- *Avg number of unique pos tags* The average number of unique POS tags per sentence
- *Lower case/upper case* The ratio of lower case characters to upper case characters
- *Word length* The word length distribution, number of words of length 1,2,3...8 and the average word length
- *Adjective rate* The number of adjectives divided by the number of tokens
- *Adverb rate* The number of adverbs divided by the number of tokens

We selected our feature set based on the work presented in [3], [14], [11] on using stylometric features for genre detection and on the work of Grieve et al. on stylometric analysis for authorship attribution [13]. We also introduced an emoticon distribution due to the special style elements on the Web, so-called emoticons. Emoticons are a series of punctuations and characters which many browsers and tools interpret as different smileys.

B. Lexical Features

Aside to the stylometric features, we applied the following common lexical features to both classification tasks:

- *Unigrams* Feature space with distinct words which occur in the text
- *Bigrams* Feature space with groups of two words which occur sequentially within a sentence
- *Trigrams* Feature space with groups of three words which occur sequentially within a sentence
- *Stems* Feature space with words stemmed by the porter stemming algorithm implemented within the snowball stemming framework[1]
- *Nouns* Feature space with all nouns annotated by the OpenNLP library[2]
- *Verbs* Feature space with all verbs
- *Adjectives* Feature space with all adjectives and adverbs
- *Leading graphem* Feature space with character ngrams of size three leading any token in the text
- *Trailing graphem* Feature space with character ngrams of size three trailing any token in the text
- *Pers.pronouns* Feature space with all distinct personal pronouns occurred in the text

IV. Experiments

Due to the lack of a standard corpus related to our work, we annotated a subset of the TREC Blogs08 Dataset[3]. We manually annotated 83 blogs with total number of 12844

[1] http://snowball.tartarus.org/

[2] http://opennlp.sourceforge.net/

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

distinct blog entries in English into "news related", and "other" blogs and "neutral", and "emotional".

	News Related	Other
blog level	29%	71%
entry level	30%	70%

(a) News vs. Rest

	Emotional	Neutral
blog level	52%	48%
entry level	40%	60%

(b) Emotion Classification Task

Table I
CORPUS DISTRIBUTIONS

TableI shows the percentage of each class for both tasks and levels. The blogs for the corpus were selected at random from the 1.3 million blogs available in the TREC Blogs08 Corpus.

To identify the most valuable stylometric features, we calculated the mutual information. The mutual information is a measure for arbitrary dependency between random variables and has been used extensively in the literature for feature selection [15], [16] because mutual information is invariant under linear transformations and takes into account the whole dependency structure of the random variables.

Figure 1 shows the mutual information of the top 20 stylometric features for both the news versus rest task and the emotion classification task. The mutual information reveals for the news versus rest task that the *adjectives+adverbs/token* feature is the most correlated with about 0.36. Also, the word length, the sentence complexity (number of words per sentence), and the *adjectives/token* are highly correlated features for the news versus rest task.

For the emotion classification task, apparently the number of used adverbs is much more relevant since the *adverbs/token* feature is the most correlated feature. An interpretation might be that an individual, who writes very emotionally often uses many adverbs whereas a neutral writer uses adverbs rarely which is also indicated in [6]. The *lower case/upper case* feature is the second most correlated feature for this task, and similar as in the news versus rest task, features denoting the word/sentence complexity (*Chars in sentences* and *Words in sentences*) are highly correlated.

For complex feature spaces based on terms and tokens, such as the term space, we evaluated state of the art classifiers and report the accuracy of these classifiers on our two tasks. All results are retrieved using a 10-fold cross validation. The result of this matrix evaluation based on three classifiers, two different tasks, two different annotation levels and 10 different feature spaces is shown in Figure 2 and 3.

From the 60 experiments on the news versus rest task, we made three observations: Firstly, it makes a difference on which level the classification is performed. For instance, the accuracy of the best performing classification experiment on blog level is 81.3% (KNN on stems) and the best accuracy on entry level is 91.2% (LibLinear on stems). The summation of all blog entries into a single blog document performs about 10% worse. To assess the overall blog genre, it is therefore necessary to evaluate the genre on entry level and then to extrapolate this to the blog level. Secondly, in Figure 2a one can clearly see that the differences in between the accuracy on the different feature spaces are less than for the entry level (Figure 2b). We assume that this "smoothing' effect is an artifact of the merging process. The merging process leads to an averaged feature space and the signal to noise ratio becomes smaller. Thirdly, the assumption that the adjective feature space is discriminative is not verifiable on these experiments - as one might have expected from the statistical feature analysis. This is due to the fact that even though the ratio of adjectives are correlated for the news versus rest task, the adjective feature space is not. This means that the ratio is discriminative, which can be interpreted so that news authors use more adjectives overall, but obviously they do not use different ones.

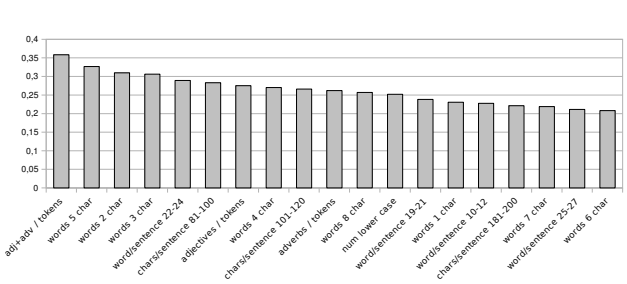
The 60 experiments on the emotion classification task confirmed the observations from the news versus rest task. Again, there is a difference on which level the experiments are conducted (blog level LibLinear on graphems 83.0 versus entry level LibLinear on stems 91.4). Also, on this task the third observation from above is confirmed: the statistical analysis from above reveals the highest correlation of the adverb rate with the labeling. But the accuracy on the adjective and adverb feature space is again about 15% worse as on the best feature space.

Algorithm	Train(s)	StdDev.	Test(s)	StdDev.
CFC	5.494	1.061	0.037	0.002
KNN	0.034	0.000	63.448	1.078
LibLinear	38.089	1.411	0.036	0.002

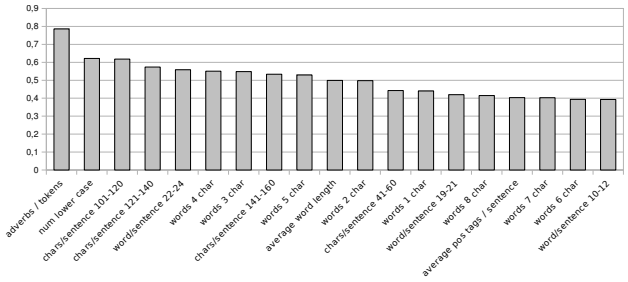
Table II
TRAIN AND TEST SPEED FOR TRIGRAMS

From an algorithmic point of view, all 120 experiments showed that the centroid based CFC algorithm performs better as the size of the feature space grows. CFC performs slightly worse on the unigram space and similar to LibLinear on the bigram and trigram feature spaces. Note that the unigram feature space has about 82k dimension, the bigram space about 680k dimensions, and the trigram space 1.42 million. Consequently, the advantages of the CFC algorithm in extremely high dimensional spaces are two-fold, firstly the algorithm performs better the more dimensions are in the feature space and secondly, the algorithm is extremely fast in terms of training and classification, see Table II. As shown in this table, the CFC training is about 10 times faster than the LibLinear training phase and the classification phase outperforms KNN by a factor of 200.

To compare the performance of stylometric features and

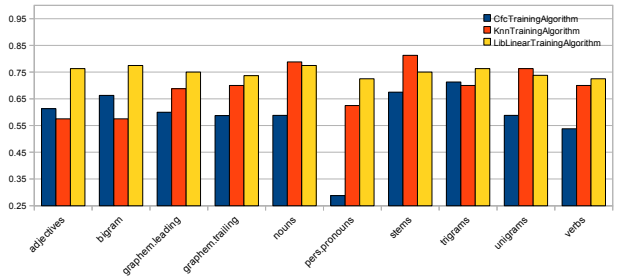


(a) News versus Rest Task

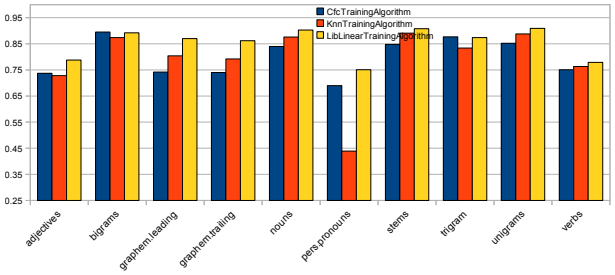


(b) Emotion Classification Task

Figure 1. Mutual Information

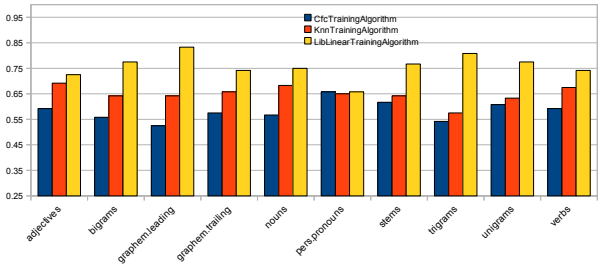


(a) Blog Level

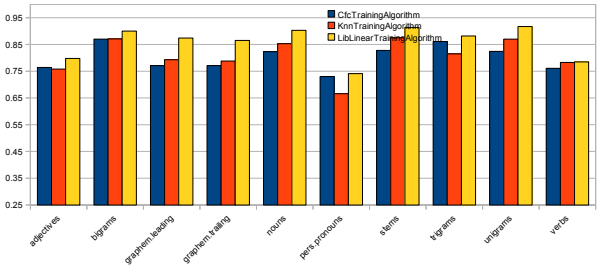


(b) Entries Level

Figure 2. News vs. Rest: Classification Accuracy



(a) Blog Level



(b) Entries Level

Figure 3. Emotion Classification Task: Classification Accuracy

lexical features, we evaluated the dense stylometric features with the same classifiers as well as with other state of the art classifiers for dense data. We used the Mallet[4] implementation of a Naive Bayes classifier, with and without boosting (AdaBoost) as well as a C45 decision tree with and without boosting (AdaBoost). Due to the fact that a linear kernel is often not the optimum for dense data, we also applied the LibSVM[17] on the classification problem with the more general RBF kernel, as suggested by the LibSVM user guide. We further did a grid search for the SVM parameters C ($2^{-3} - 2^5$) and γ ($2^{-6} - 2^3$). In this section, we only report the best performing parameter set ($C = 2^5$, $\gamma = 2^1$). Experiments with 10-fold cross validation revealed that the stylometric features are no match for both tasks compared to the high dimensional feature space classifiers. The results for this experiments are shown in Table III.

The best stylometric feature based classification resulted in a significantly lower accuracy than achieved with the lexical features. For these experiments, we took the features with the highest mutual information into account and were left with a number of distinct features of 24. Nevertheless, stylometric features are guaranteed to be topic independent and therefore their generalization capability is higher [18].

[4]http://mallet.cs.umass.edu/

Algorithm	News	Emotionality
CFC	0.69	0.73
LibLinear	0.72	0.78
k-NN10	0.74	0.78
LibSvm	0.69	0.73
NB	0.70	0.76
NB+AdaBoost	0.70	0.77
C45	0.72	0.78
C45+AdaBoost	0.72	0.76

Table III
CLASSIFICATION ACCURACY OF STYLOMETRIC FEATURES.

V. Conclusions

In this work, we investigated lexical and stylometric classification features to determine the best performing features and classifiers to assign blogs to the category news and to assess the emotionality in the resulting news related blogs. This enables us to support blog retrieval by the genre news and the facets emotional versus neutral. In our experiments, we conducted a matrix evaluation on 20 features and three and/or eight state of the art classifiers with the following impact: Firstly, the news genre and the emotion must be assessed on a per entry level. Secondly, topic independent stylometric/shallow text features can be used to perform both tasks albeit with a lower accuracy than the bag of words approach but with the advantage of being guaranteed topic-independent. Finally, we were able to show that the CFC algorithm performs equally good as SVMs in high dimensional spaces (greater than 1 mill. dimensions), but outperforms LibLinear in terms of time consumption. For future work, we want to extend our work to much larger datasets. We also want to deeper investigate feature normalization techniques on the stylometric features. We assume that this would improve the performance of the SVM. Also, we plan to study the news versus rest task as a one class problem.

Acknowledgment

The Know-Center GmbH Graz 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.

References

[1] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the web," *Commun. ACM*, vol. 51, no. 12, pp. 68–74, 2008.

[2] C. S. Lim, K. J. Lee, and G. C. Kim, "Automatic genre detection of web documents," *Lecture Notes in Computer Science*, vol. 3248, pp. 310–319, 2005.

[3] B. Kessler, G. Numberg, and H. Hinrich Schütze, "Automatic detection of text genre," in *ACL-35: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 1997, pp. 32–38.

[4] C. S. Lim, K. J. Lee, and G. C. Kim, "Multiple sets of features for automatic genre classification of web documents," *Information Processing & Management*, vol. 41, pp. 1263–1276, 2005.

[5] K. H.-Y. Lin, C. Yang, and H.-H. Chen, "Emotion classification of online news articles from the reader's perspective," in *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 220–226.

[6] P. Chesley, B. V. L., Xu, and R. K. Srihari, "Using verbs and adjectives to automatically classify blog sentiment," Computational Approaches to Analyzing Weblogs: Papers from the 2006 Spring Symposium, N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin, Eds. AAAI Press, Menlo Park, CA, 27–29, Tech. Rep., 2006.

[7] E. Lex, M. Granitzer, M. Muhr, and A. Juffinger, "Stylometric features for emotion level classification in news related blogs," in *Proceedings of the 9th RIAO Conference (RIAO 2010)*, 2010.

[8] R. Fan, K. Chang, C. Hsieh, X. W. C., and Lin, "Liblinear: A library for large linear classification." *Journal of Machine Learning Research*, 2008.

[9] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.

[10] 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: ACM, 2009.

[11] J. Karlgren and D. Cutting, "Recognizing text genres with simple metrics using discriminant analysis," in *Proceedings of the 15th conference on Computational linguistics*, 1994, pp. 1071–1075.

[12] W. Sanders, *Linguistische Stilistik. Grundzge der Stylanalyse sprachlische Kommunikation*. Kleine Vandenhoeck-Reihe, Gttingen., 1977.

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

[14] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in *Proceedings of the 18th conference on Computational linguistics*, 2000, pp. 808–814.

[15] A. Blum, P., and Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, pp. 245–271, 1997.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, 2003.

[17] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[18] E. Lex, A. Juffinger, and M. Granitzer, "Objectivity classification in online media," in *Proceedings of the 21th ACM Conference on Hypertext and Hypermedia*, 2010.