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

Elisabeth Lex

Know Center Graz

30.08.2010

http://www.**know-center**.at

Agenda

- Introduction
- Aim of this work and Approach
- Features
- Experiments and Results
- Lessons Learned
- Future Work

Introduction

- A lot of content on the Web and in the blogosphere
- New challenges for search engines

→"finding a needle in a haystack to a process of being presented with a variety of needles and choosing the one you want" [Etzioni2008].

→ providing different facets for the users' different information needs is crucial

Agenda

- Introduction
- Aim of this work and Approach
- Features
- Experiments and Results
- Lessons Learned
- Future Work

Aim of this work

- Support blog retrieval with genre information and facets
- Assign blogs to the news genre and to assess a facet denoting the emotionality in the blogs
- Study which features and algorithms serve best

Approach

- Lexical versus stylometric/shallow text features to classify blogs into
  - News related blogs versus rest
  - Emotional versus neutral

Simple features, easy to extract

Approach

- Lexical versus stylometric/shallow text features to classify blogs into
  - News related blogs versus rest
  - Emotional versus neutral
- Feature study on lexical features with 3 SOTA classifiers
- Analysis of statistical properties of stylometric features with Mutual Information (MI)

Approach

- Supervised classification strategy:
- Binary Problems: News versus Rest, Emotional versus Neutral
  - SVM (LibLinear, LibSVM)
  - K-NN (k=10, both cosine as well as Euclidean)
  - Class Feature Centroid (CFC) (b=1.1)

Agenda

- Introduction
- Aim of this work and Approach
- Features
- Experiments and Results
- Lessons Learned
- Future Work

Features – Lexical Features

- Common Bag of Words features:
  - Unigrams, Bigrams, Trigrams
  - Stems (Porter Stemming Algorithm)
  - Nouns, Verbs, Adjectives (OpenNLP)
  - Leading and Trailing Graphems
  - Personal Pronouns

We create one feature space for each type of feature!

Features – Stylometric Features

- Stylistic variations depend on author, genre, context, characteristics of indented audience [Sanders1977]
  - Punctuation, Emoticons
  - Words in sentences, Avg words / sentences
  - Chars in sentences, Avg chars / sentences
  - Noun+verb sentences (complete sentences)
  - Avg number of unique pos tags
  - Lower case/upper case
  - Word length
  - Adjective rate and adverb rate

Features – Stylometric Features

- Stylistic variations depend on author, genre, context, characteristics of intented audience [Sanders1977]
  - Punctuation, Emoticons
  - Words in sentences, Avg words / sentences
  - Chars in sentences, Avg chars / sentences
  - Noun+verb sentences (complete sentences)
  - Avg number of unique pos tags
  - Lower case/upper case
  - Word length
  - Adjective rate and adverb rate

Focus on topic independence: Text classifiers can easily overfit to topics due to natural correlation between topics and genres

Agenda

- Introduction
- Aim of this work and Approach
- Features
- Experiments and Results
- Lessons Learned
- Future Work

Experiments - Dataset

- Manually annotated subset of Blogs08 TREC dataset: 83 annotated blogs (12844 Blog entries)

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

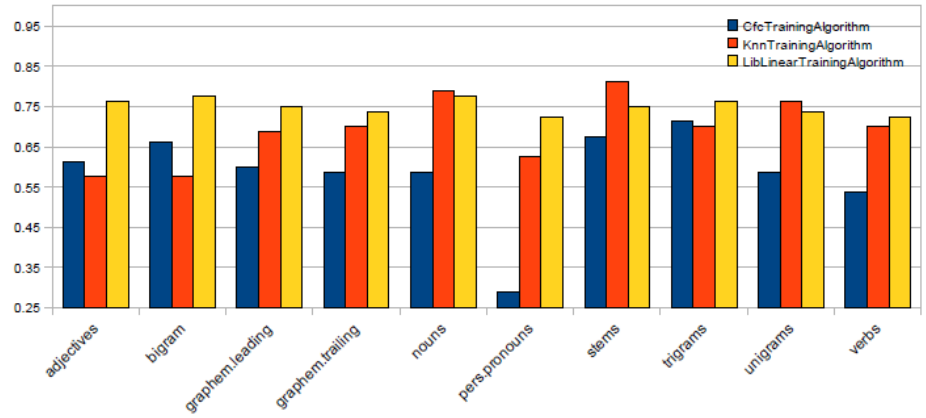
(a) News vs. Rest

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

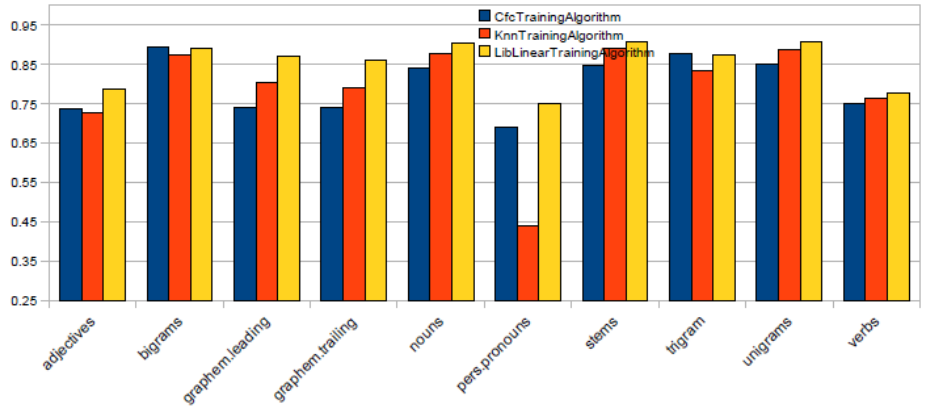
(b) Emotion Classification Task

Table I
CORPUS DISTRIBUTIONS

Results – Lexical Features: News versus Rest



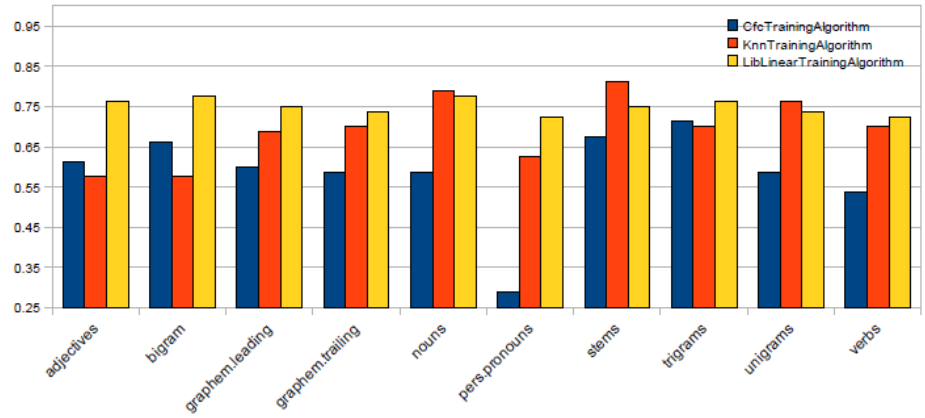
(a) Blog Level



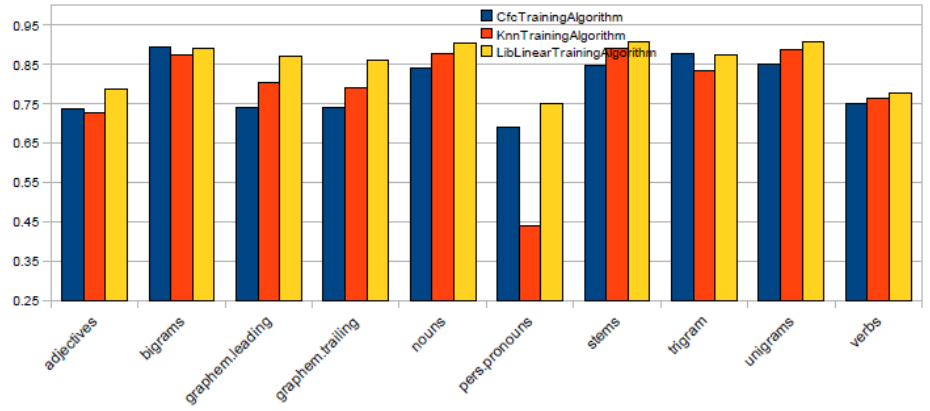
(b) Entries Level

Figure 2. News vs. Rest: Classification Accuracy

Results – Lexical Features: News versus Rest



(a) Blog Level



(b) Entries Level

Figure 2. News vs. Rest: Classification Accuracy

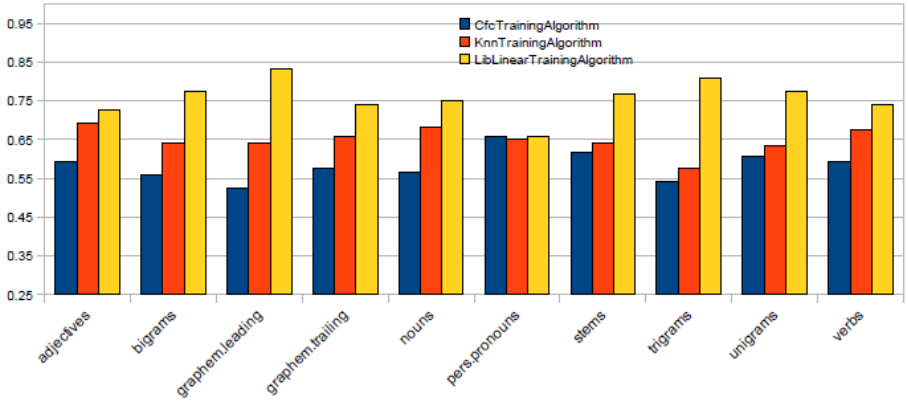
→it makes a difference on which level the classification is performed!

-Best result on entry level: 91.2% (LibLin on stems)

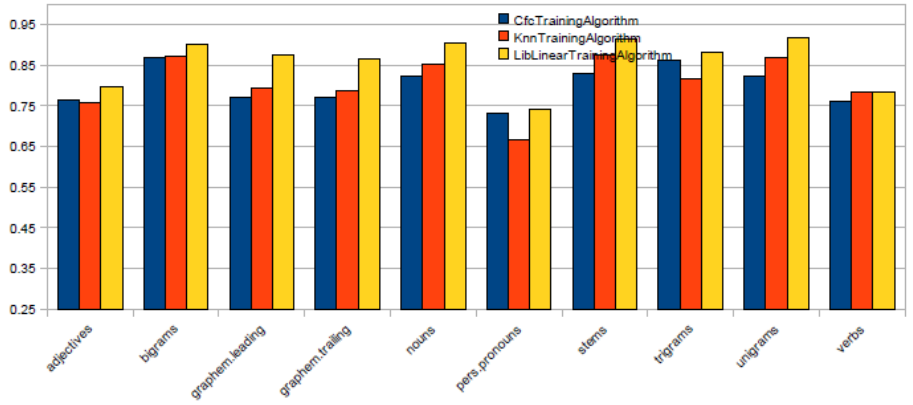
- Best result on blog level: 81.3% (kNN on stems)

→ Assess genre on entry level and extrapolate to the blog level

Results – Lexical Features: Emotionality



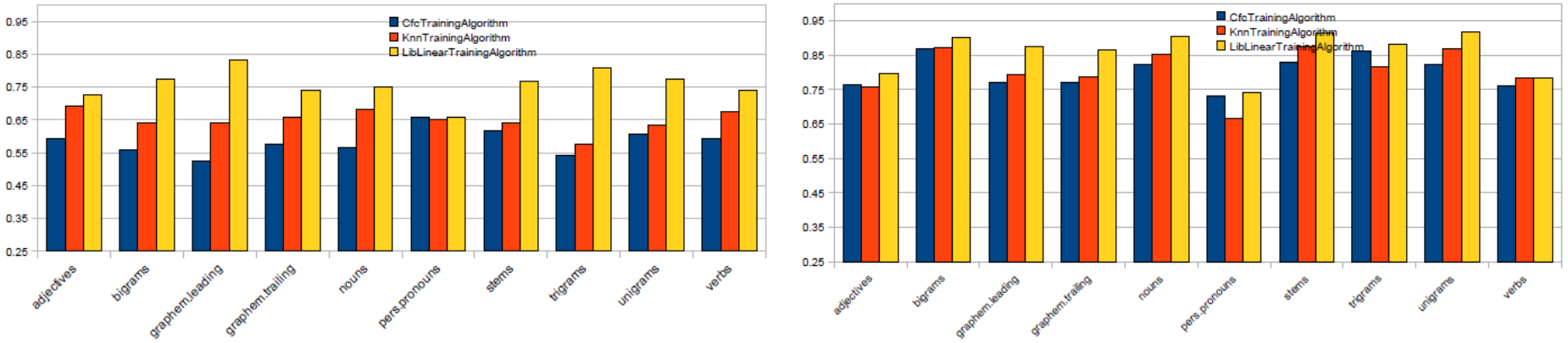
(a) Blog Level



(b) Entries Level

Figure 3. Emotion Classification Task: Classification Accuracy

Results – Lexical Features: Emotionality



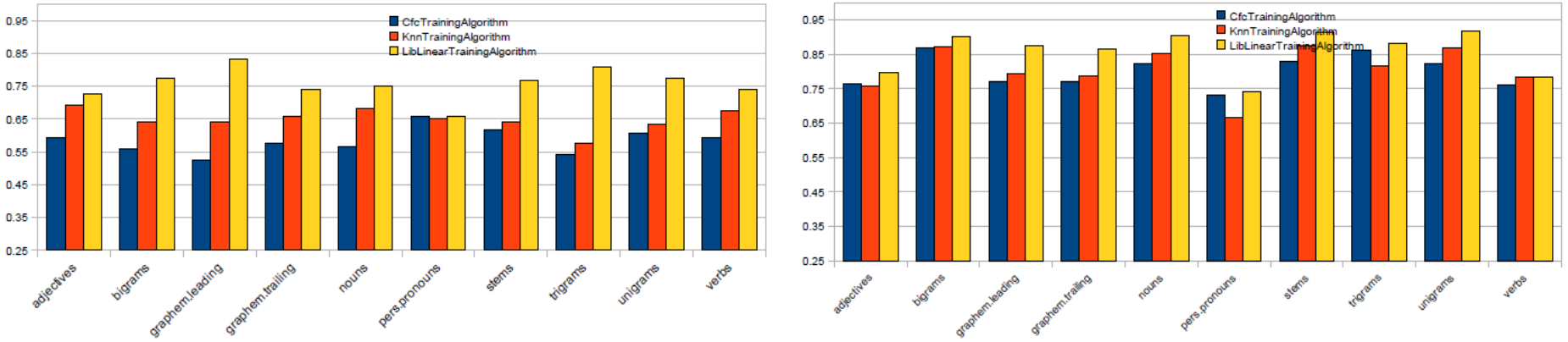
(a) Blog Level (b) Entries Level

Figure 3. Emotion Classification Task: Classification Accuracy

→Best result: LibLin on stems: 91.4% (entry level)

→Good results with stems, nouns → typically topic oriented features

Results – Lexical Features: Emotionality



(a) Blog Level (b) Entries Level

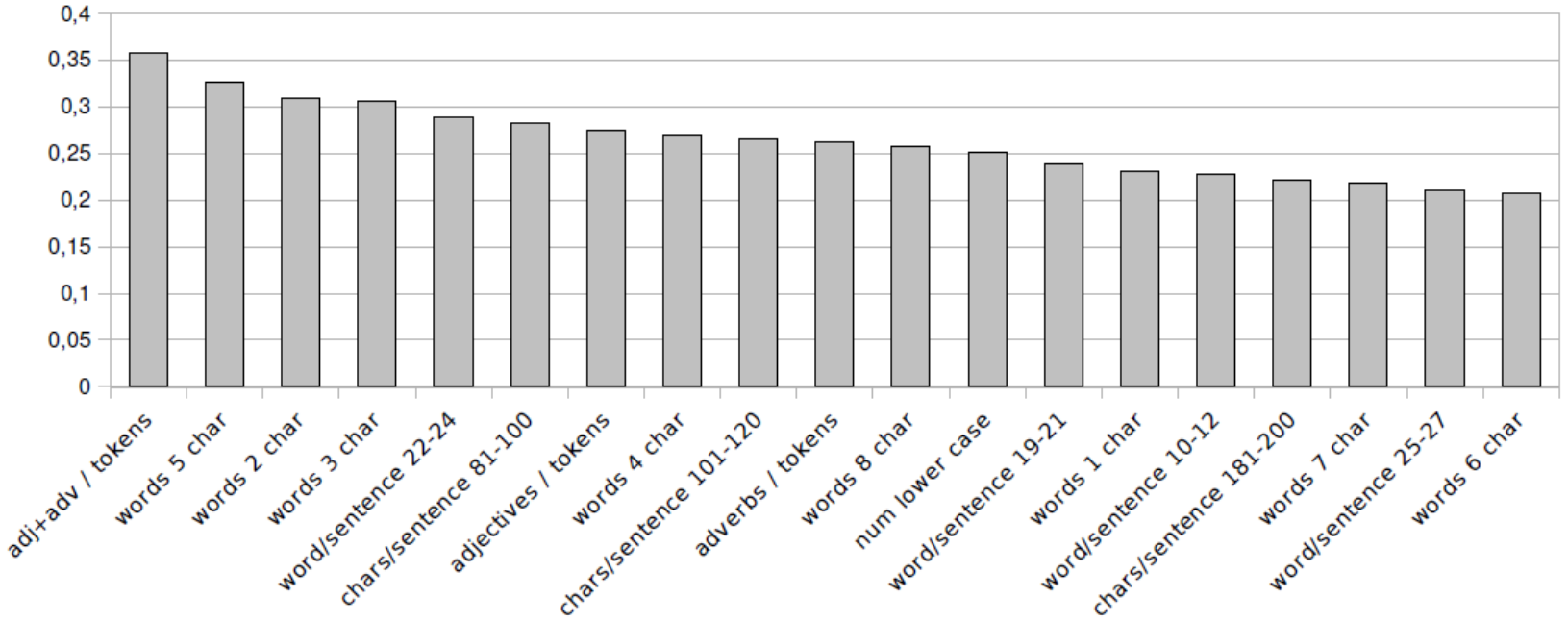
Figure 3. Emotion Classification Task: Classification Accuracy

→Best result: LibLin on stems: 91.4% (entry level)

→Good results with stems, nouns → typically topic-oriented features

Focus on topic independence: Text classifiers can easily overfit to topics due to natural correlation between topics and genres

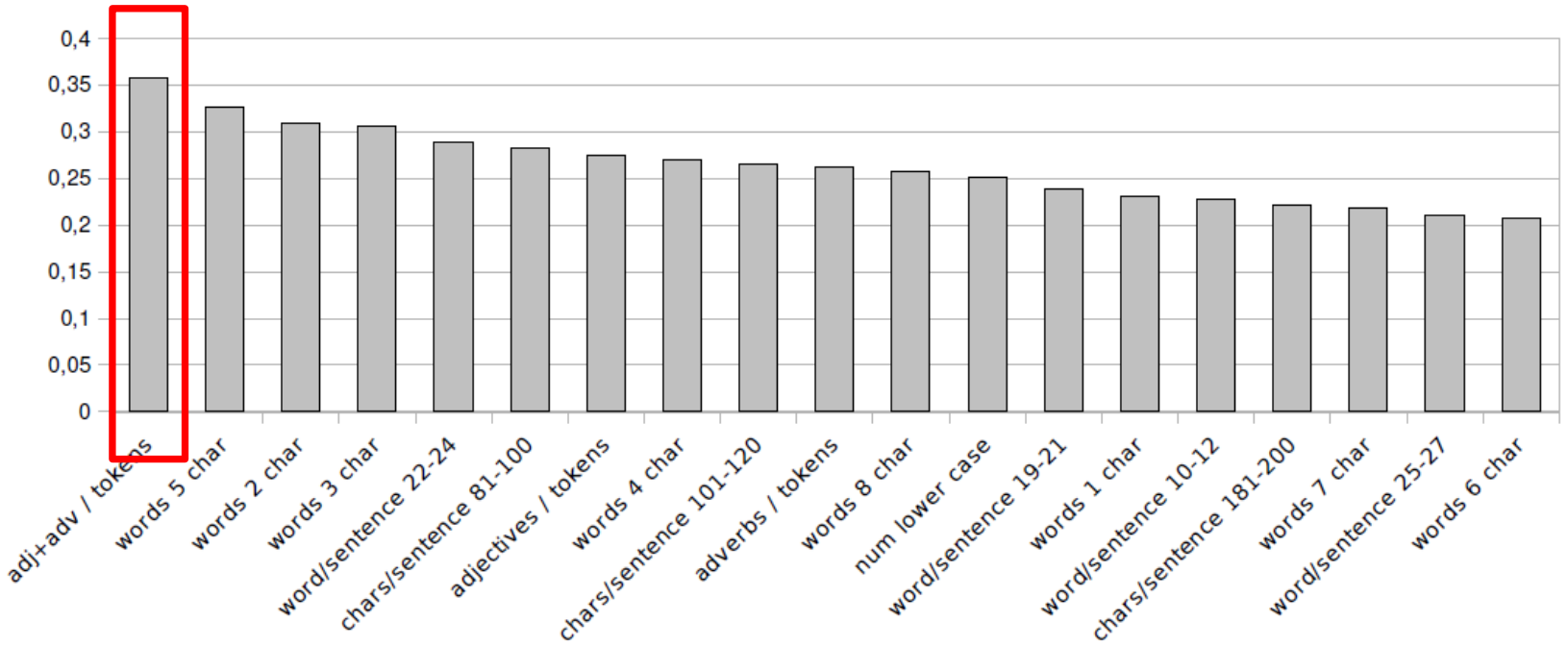
Experiments – Mutual Information for Stylometric Features for News versus Rest



(a) News versus Rest Task

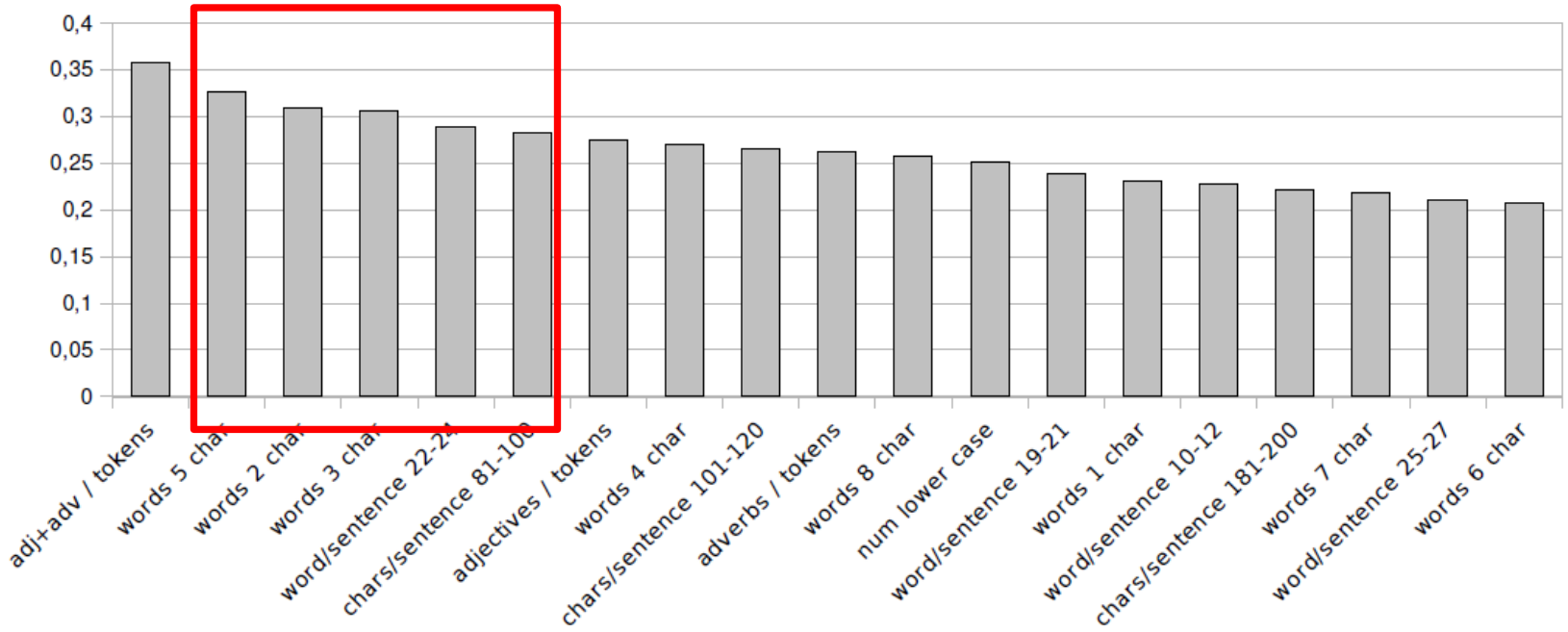
I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", Journal of Machine Learning Research, 2003

Experiments – Mutual Information for Stylometric Features for News versus Rest



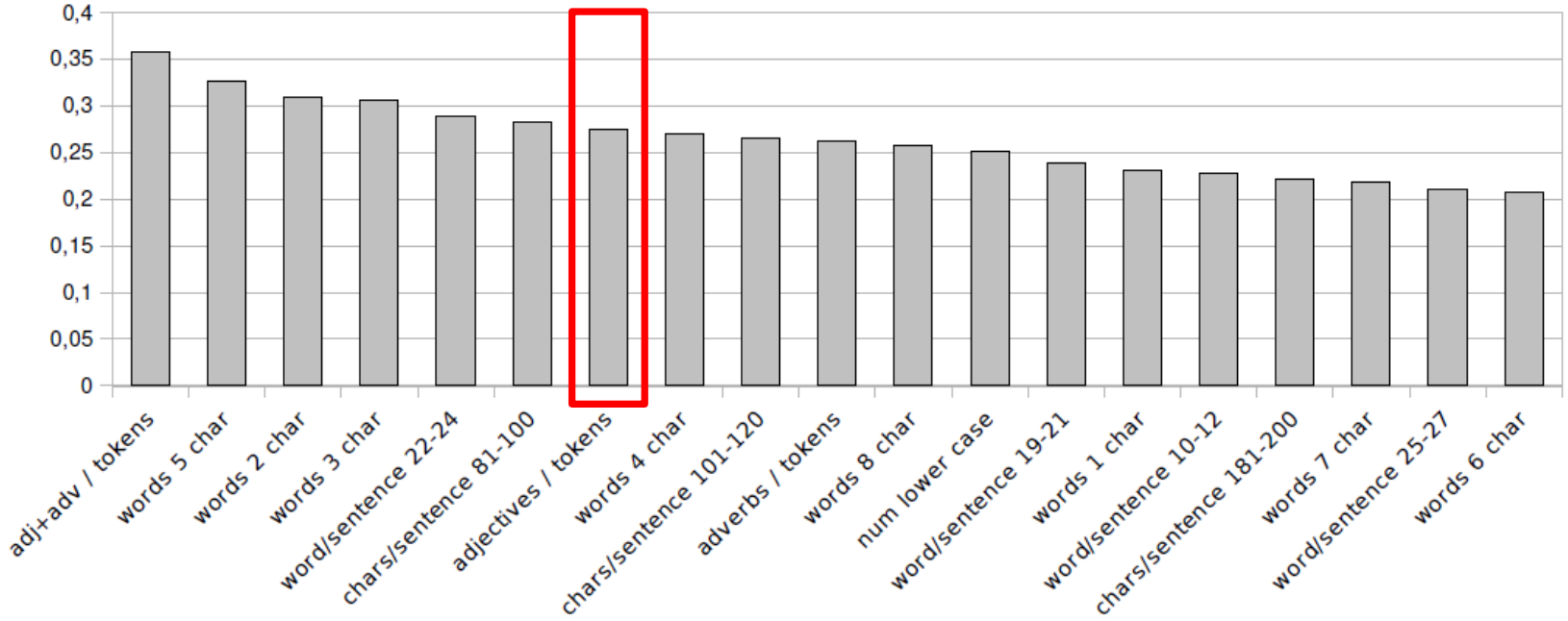
(a) News versus Rest Task

Experiments – Mutual Information for Stylometric Features for News versus Rest



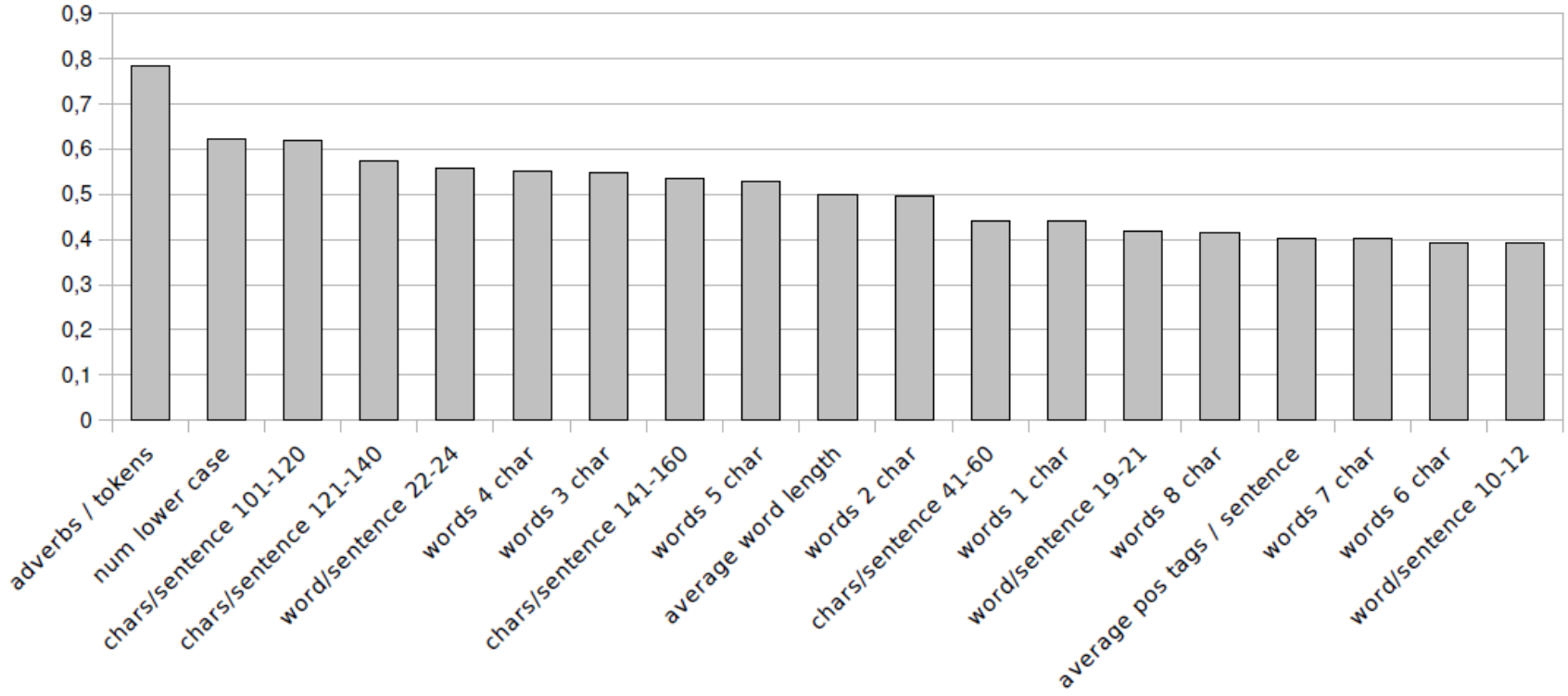
(a) News versus Rest Task

Experiments – Mutual Information for Stylometric Features for News versus Rest



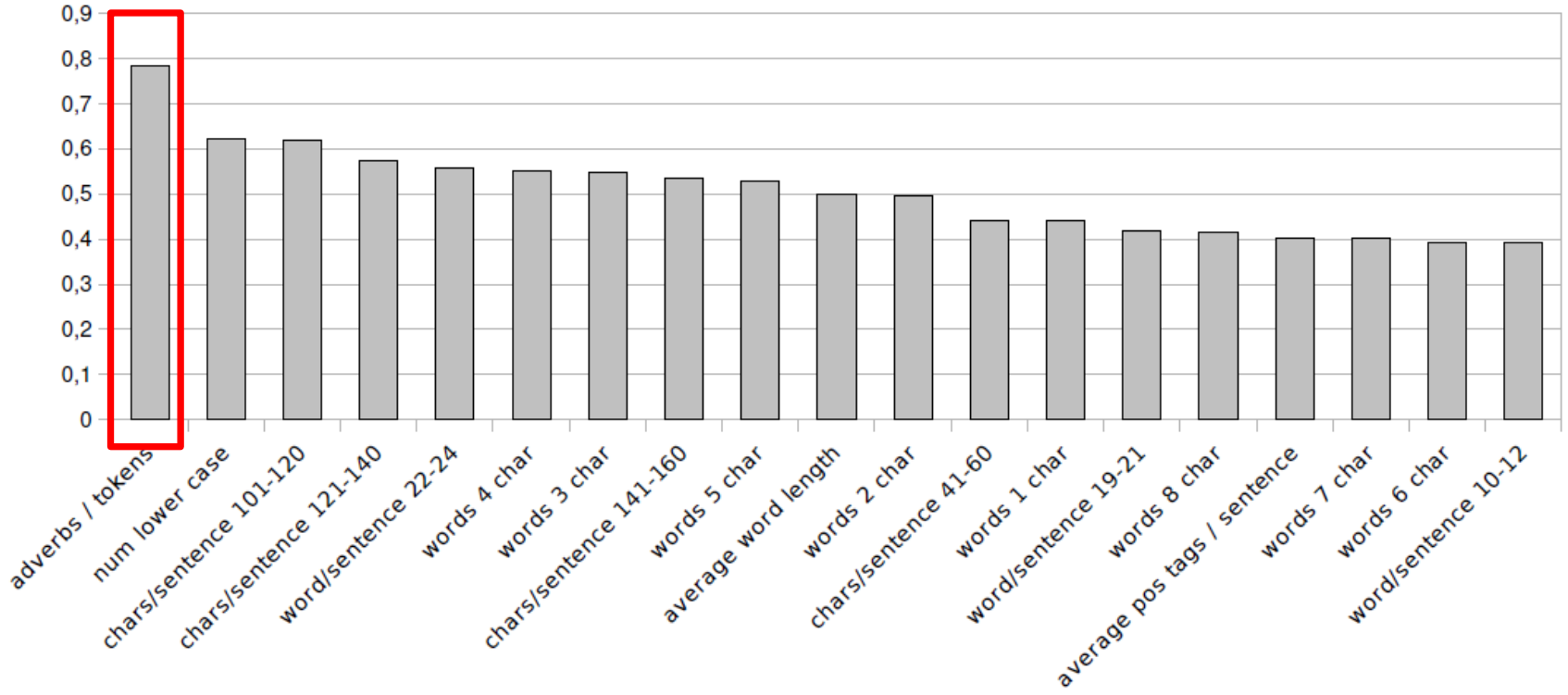
(a) News versus Rest Task

Experiments – Mutual Information for Stylometric Features for Emotion Task



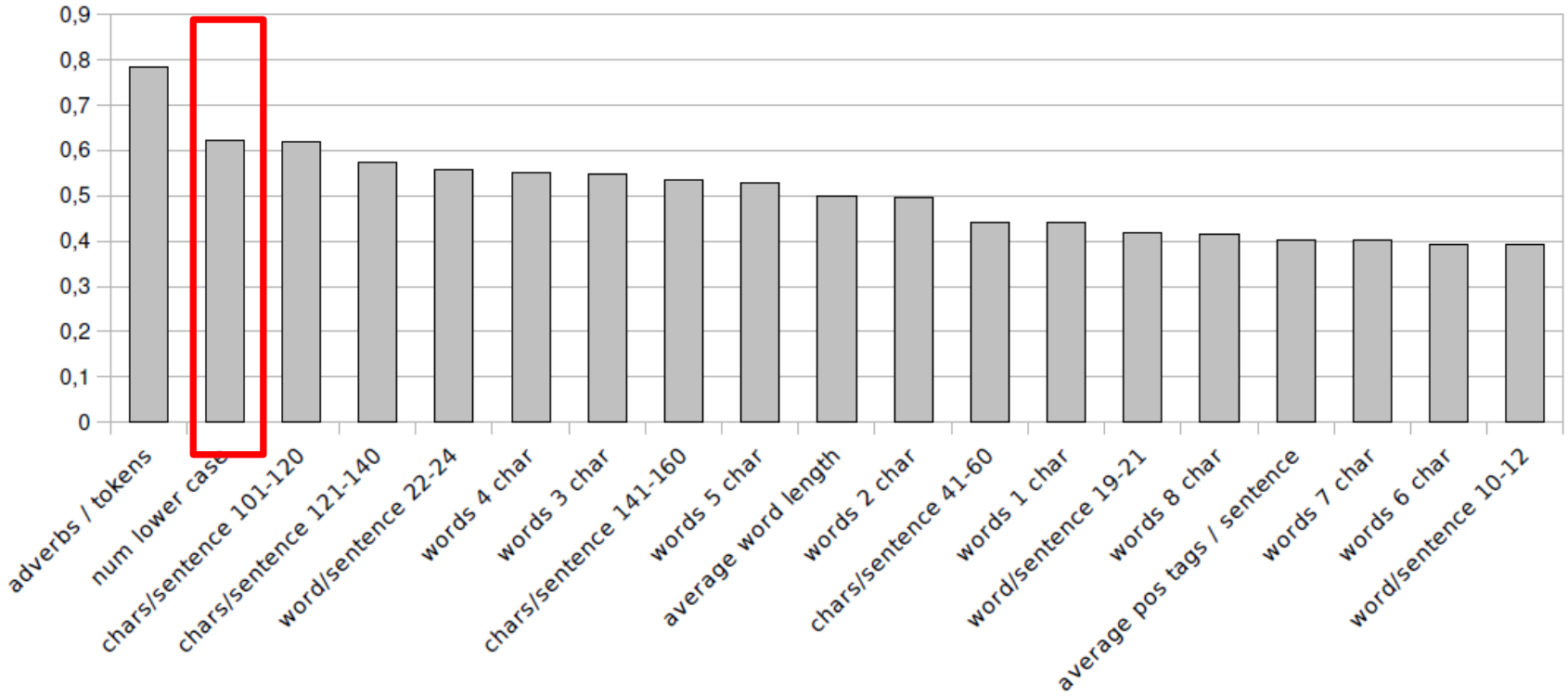
(b) Emotion Classification Task

Experiments – Mutual Information for Stylometric Features for Emotion Task



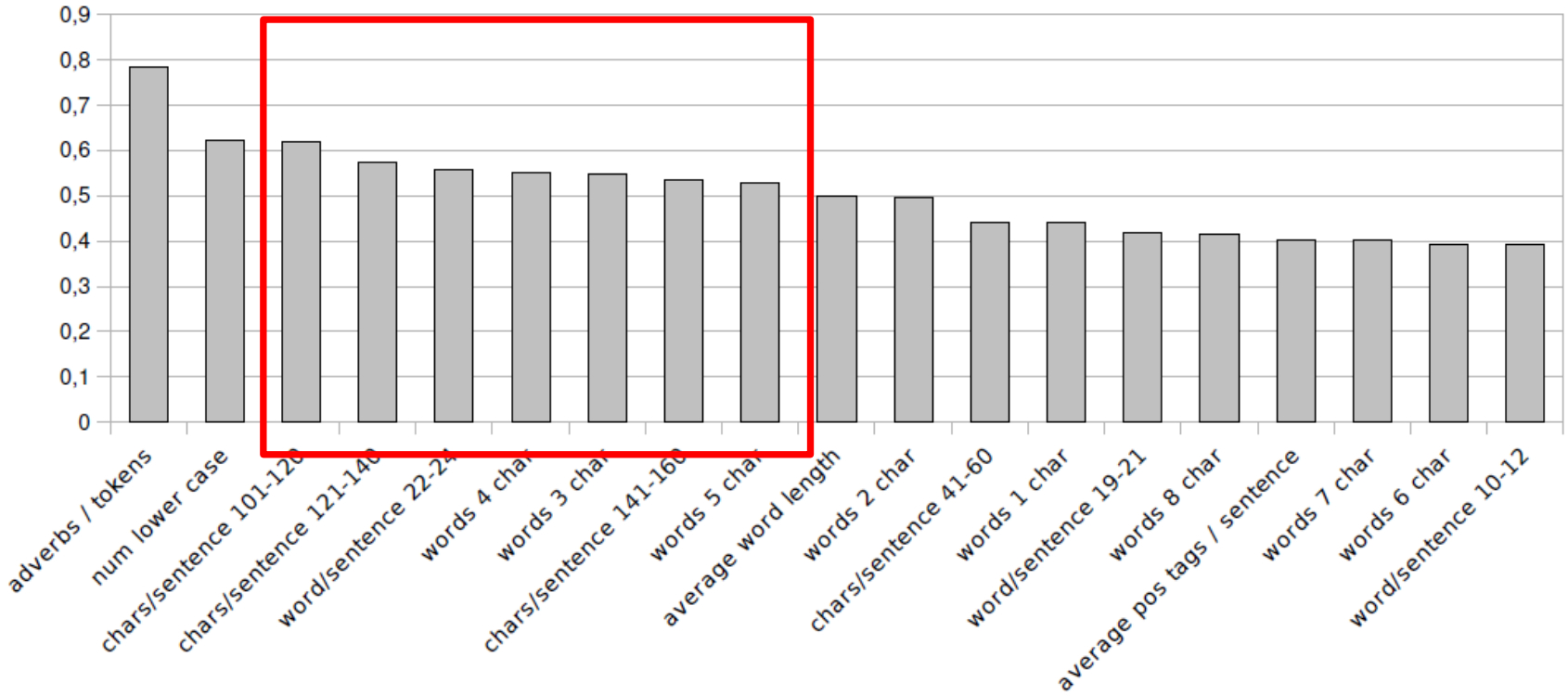
(b) Emotion Classification Task

Experiments – Mutual Information for Stylometric Features for Emotion Task



(b) Emotion Classification Task

Experiments – Mutual Information for Stylometric Features for Emotion Task



(b) Emotion Classification Task

Results – Stylometric Features

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.

→ Only the best features, according to their Mutual Information
→ No match for lexical features
→ But: Topic independent – generalize better [Lex2010]

Results: Algorithmic Point of View

- CFC performs best when size of feature space grows and feature space is sparse
  - Unigram space: 82k dimensions, LibLin wins
  - Bigram space: 680k dimensions, LibLin and CFC same
  - Trigram space: 1.42 Mio dimensions, CFC wins
- CFC fast in terms of training and testing

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

- In dense stylometric feature space – kNN10 wins

Agenda

- Introduction
- Aim of this work and Approach
- Features
- Experiments and Results
- Lessons Learned
- Future Work

Lessons Learned

- The news genre and the emotionality must be assessed on a per entry level
- Topic independent stylometric text features can be used to perform the emotion classification task

  → However, their accuracy is lower

  → But: Topic independent
- Classifiers trained on lexical features perform consistently better than classifiers trained on the best stylometric features
- The CFC algorithm performs equally good as SVM in high dimensional spaces (> 1 Mio dimensions), but outperforms LibLinear in terms of time consumption

Agenda

- Introduction
- Aim of this work and Approach
- Features
- Experiments and Results
- Lessons Learned
- Future Work

Future Work

- Study the genre classification problem as One Class problem
- Combine the best lexical features in one feature space
- Extension to much larger datasets

Please find our annotations of the Blogs08 TREC dataset on our Website:

www.know-center.at/forschung/knowledge_relationship_discovery/downloads_demos/annotated_blog_corpus_facet_annotations_on_the_trec_blogs08_test_collection

Thank you for your attention!
Questions?

References

[Sanders1977] W. Sanders, Linguistische Stilistik. Grundzüge der Styleanalyse sprachliche Kommunikation. Kleine Vandenhoeck-Reihe, Göttingen., 1977.

[Lex2010] E.Lex, A. Juffinger, and M. Granitzer, Objectivity classification in Online Media. ACM HT 2010.

[Grieve2007] J. Grieve, Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 2007.