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
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)

<table>
<thead>
<tr>
<th></th>
<th>News Related</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>blog level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entry level</td>
<td>29%</td>
<td>71%</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>70%</td>
</tr>
</tbody>
</table>

(a) News vs. Rest

<table>
<thead>
<tr>
<th></th>
<th>Emotional</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>blog level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>entry level</td>
<td>52%</td>
<td>48%</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>60%</td>
</tr>
</tbody>
</table>

(b) Emotion Classification Task

Table I
CORPUS DISTRIBUTIONS

http://www.know-center.at
Results – Lexical Features: News versus Rest

Figure 2. News vs. Rest: Classification Accuracy
Results – Lexical Features: News versus Rest

→ 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

Figure 2. News vs. Rest: Classification Accuracy
Results – Lexical Features: Emotionality

Figure 3. Emotion Classification Task: Classification Accuracy

(a) Blog Level

(b) Entries Level
Results – Lexical Features: Emotionality

- Best result: LibLin on stems: 91.4% (entry level)
- Good results with stems, nouns → typically topic oriented features
Results – Lexical Features: Emotionality

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

http://www.know-center.at
Experiments – Mutual Information for Stylometric Features for News versus Rest

(a) News versus Rest Task


http://www.know-center.at
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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>News</th>
<th>Emotionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFC</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>LibLinear</td>
<td>0.72</td>
<td>0.78</td>
</tr>
<tr>
<td>k-NN10</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td>LibSvm</td>
<td>0.69</td>
<td>0.73</td>
</tr>
<tr>
<td>NB</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>NB+AdaBoost</td>
<td>0.70</td>
<td>0.77</td>
</tr>
<tr>
<td>C45</td>
<td>0.72</td>
<td>0.78</td>
</tr>
<tr>
<td>C45+AdaBoost</td>
<td>0.72</td>
<td>0.76</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Train(s)</th>
<th>StdDev.</th>
<th>Test(s)</th>
<th>StdDev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFC</td>
<td>5.494</td>
<td>1.061</td>
<td>0.037</td>
<td>0.002</td>
</tr>
<tr>
<td>KNN</td>
<td>0.034</td>
<td>0.000</td>
<td>63.448</td>
<td>1.078</td>
</tr>
<tr>
<td>LibLinear</td>
<td>38.089</td>
<td>1.411</td>
<td>0.036</td>
<td>0.002</td>
</tr>
</tbody>
</table>

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

