Naïve Bayesian Classifiers for Ares News Articles

Derek Singer (derekcs@rice.edu), Dr. Devika Subramanian (devika@rice.edu)

Problem: Overlapping Categories

• Want: Articles relevant to war and international relations
• Don’t want: Articles only about irrelevant matters such as economy and local/national politics

Example Article: Congress passes wheat embargo on Iran.

IRRELEVANT
RELEVANT

Belongs to more irrelevant categories than relevant but is relevant.

Problem: Overlapping Words

\[ NBF(d) = \arg \max_{c \in \{R, I\}} P(c) \sum_{w \in d} \log(P(w|c)) \]

\( C = \) Class, \( D = \) Document, \( W = \) Words (a.k.a. tokens)

Overlap of Tokens between Categories

78.87% of unique R tokens
12.29% of unique I tokens
99.1% of total R tokens
85.1% of total I tokens

Mean ratio of probabilities

\[ \frac{1}{|\text{tokens}|} \sum_{w \in \text{tokens}} \frac{P(w|R)}{P(w|I)} = 9.6094 \]

The large majority of tokens create a bias in the filter towards Relevant.

Either-Or Classification Fails

• Reuters articles 8/20/1996-8/19/1997
• 61,005 Relevant, 745,759 Irrelevant
• Using All Tokens of Length > 2, 5-fold Cross Validation

Results with Finer-Grained Categories

• Reuters articles 8/20/1996-8/19/1997
• Using All Tokens of Length > 2, 5-fold Cross Validation

<table>
<thead>
<tr>
<th># of Articles</th>
<th>MEAN</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>567,052</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>112,611</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>355</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>66,123</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>198</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>191,196</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>1,019</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>X’s Total Token Overlap w/ X</th>
<th>X’s Total Token Overlap w/ G</th>
<th>Mean Ratio of Probabilities (G : X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>49.25%</td>
<td>11.24%</td>
<td>17.93</td>
</tr>
<tr>
<td>B</td>
<td>32.82%</td>
<td>26.82%</td>
<td>2.21</td>
</tr>
<tr>
<td>C</td>
<td>19.44%</td>
<td>13.33%</td>
<td>0.20</td>
</tr>
<tr>
<td>D</td>
<td>19.64%</td>
<td>13.04%</td>
<td>0.90</td>
</tr>
<tr>
<td>E</td>
<td>11.71%</td>
<td>13.24%</td>
<td>0.12</td>
</tr>
<tr>
<td>F</td>
<td>13.94%</td>
<td>13.68%</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Fix: Finer-Grained Categories

• Relevant and Irrelevant not distinct enough categories
• Can instead classify according to what mix of general categories an article may belong to

New categories:

• \( A = \) Economics, Markets, Capital
• \( B = \) Government, Science, Entertainment, etc.
• \( C = \) War, International Relations
• \( D = A \) and \( B \)
• \( E = A \) and \( C \)
• \( G = A, B, \) and \( C \)

• Relevant now defined as “belongs to category 3, 5, 6 or 7
• Irrelevant now defined as “belongs to category 1, 2, or 4
• Allows for larger margin of error in classification (e.g. a category 6 article can be correctly classified as category 3, 5, or 7 as well as 6)

Example Article: Congress passes wheat embargo on Iran.

Fix: Computed Features

• How many countries mentioned?
• How many of those countries are in the Middle East?
• How long is the article?
• How many numbers appear?
• Do the people mentioned in the article represent a nation? (e.g. President Bush ⇒ USA)

Example article: Weather reports for the Middle East

• Israel, Iraq, Iran mentioned
• 10+ countries mentioned
• 10+ numbers listed

LIKELY TO BE IRRELEVANT

Future Work

• Implement computed features
• Try more sophisticated classifying algorithms (OKAPI, Support Vector Machines, Transformed Weight Complemented Naïve Bayesian)
• Reduce bias of small samples of relevant categories by bagging either-or classifiers for every pair of one relevant and one irrelevant finer-grained categories

Conclusions

• Merging all articles into two categories creates high overlap in subject matter and tokens
• High token overlap biases the filter towards one category
• Splitting categories into more natural classifications helps reduce token overlap and filter bias
• Recall, accuracy, and especially precision are improved by splitting the categories

Bibliography

• On Predicting Rare Classes with SVM Ensembles in Scene Classification. Rong Yan, Yan Liu, Rong Jin, Alex Hauptmann. ICASSP. 6-10 April 2003.