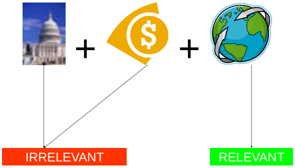
Naïve Bayesian Classifiers for Ares News Articles

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.



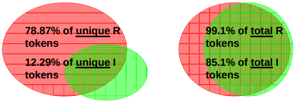
Belongs to more irrelevant categories than relevant but is relevant.

Problem: Overlapping Words

$$NBF(d) = \arg\max_{c \in \{R,I\}} P(c) \sum_{w \in d} \log(P(w \mid c))$$

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

Overlap of Tokens between Categories



$$\textbf{Mean ratio of probabilities} = \frac{1}{|tokens|} \sum_{w \in tokens} \frac{P(w \mid R)}{P(w \mid I)} = 9.6094$$

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

	MEAN	STD
REC	98.03%	0.12%
PREC	45.32%	0.36%
ACC	90.91%	0.01%

Recall = % of relevant articles classified as relevant

Precision = % of articles classified as relevant that are relevant

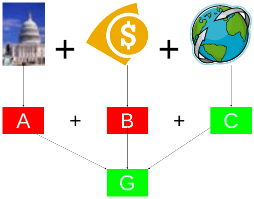
Accuracy = % of articles classified correctly

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
  - F = B 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

Results with Finer-Grained Categories

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

	# of Articles
A	567,052
B	112,611
C	353
D	66,123
E	200
F	50,256
G	10,196

	MEAN	STD
REC	95.34%	0.19%
PREC	72.16%	0.49%
ACC	96.86%	0.01%

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

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**

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

Bibliography

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

• **Tackling the Poor Assumptions of Naive Bayes Text Classifiers**. Jason D.M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger. ICML 2003.