Sentiment Analysis

• Today
  – Background and motivation
  – Sentiment categorization
    • Text categorization

Subjectivity vs. Sentiment

• **Subjective** sentences express *private states*, i.e. internal mental or emotional states
  – speculations, beliefs, emotions, evaluations,
    goals, opinions, judgments, …
    • (1) Jill said, "I hate Bill."
    • (2) Jack *thought* he won the race.
    • (3) Judy *hoped* her presentation would go well.

Subjectivity vs. Sentiment

• **Sentiment** expressions are a type of subjective expression
  – expressions of *positive* and *negative* emotions, judgments, evaluations, …
    • (1) Jill said, "I hate Bill."
    • (2) Jack *thought* he won the race.
    • (3) Judy *hoped* her presentation would go well.

Queries of a Subjective Nature

- How have **business** views towards **global climate change** varied over the past decade?
- How have **consumers and businesses** responded to the release of Gore’s “An Inconvenient Truth”?
- What is the reaction in **Asia** to the **Bush policy towards the Kyoto Protocol**?
- **Who** were the first people to criticize the proposed **bailout options for banks** in the current economic crisis?
- What does **Sarah Palin** think about <X>?
Sentiment Categorization

• Background and motivation
  → Sentiment categorization
  → – Text categorization

Text Classification (binary)

Is the overall sentiment in the document positive? negative?

Pang et al., EMNLP 2002
Turney, ACL 2002
Turney & Littman, TOIS 2003 ...

E.D. and F. MAN TO BUY INTO HONG KONG FIRM
The U.K. Based commodity house E.D. and F. Man Ltd and Singapore’s Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo’s 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will rename the firm Man Pacific (Holdings) Ltd.

About a corporate acquisition?

Yes

No
Text Classification (multi-class)

The U.K. Based commodity house E.D. and F. Man Ltd and Singapore’s Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo’s 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will rename the firm Man Pacific (Holdings) Ltd.

Text Classification

- Assign pieces of text to predefined categories based on content
- Types of text
  - Documents (typical)
  - Paragraphs
  - Sentences
  - WWW sites
- Different types of categories
  - By topic
  - By function
  - By author
  - By style

Text Classification Applications

- Help-Desk Support
  - Who is an appropriate expert for a particular problem?
- Information Filtering Agent
  - Which news articles are interesting to a particular person?
- Relevance Feedback
  - What are other documents relevant for a particular query?
- Knowledge Management
  - Organizing a document database by semantic categories.
- Focused Crawling
  - Find all the WWW pages on a particular topic.

Why Learn Text Classifiers

- Classifying documents by hand is costly and does not scale well
  - e.g. browse all WWW pages to filter out those about job announcements
- Humans are not really good at constructing text classification rules
  - It is hard to write good queries
- Sometimes there is no expert available
  - e.g. rules for routing email
- Sometimes training data is cheap and plenty
  - e.g. existing databases
Learning Setting

Goal:
- Find a classification rule $h$ with low prediction error w.r.t. the selected loss function on new examples from distribution $P(X,Y)$

Prediction Error and Loss Function

- Loss function
  - Assigns amount of "penalty" when making a mistake
  - Zero/One-Loss:
    \[
    \Delta(h(\vec{x}), y) = \begin{cases} 
    0 & \text{if } h(\vec{x}) = y \\ 
    1 & \text{else} 
    \end{cases}
    \]

- Prediction error
  - Also generalization error or true error
  - Probability of making an error on a new example drawn from the same distribution $P(X,Y)$
  - Equivalent: Expected value of loss function

Performance

- It depends…

<table>
<thead>
<tr>
<th># training docs</th>
<th>14,704</th>
<th>10,667</th>
<th>9,610</th>
<th>9,608</th>
<th>9,608</th>
</tr>
</thead>
<tbody>
<tr>
<td># categories</td>
<td>135</td>
<td>93</td>
<td>92</td>
<td>90</td>
<td>10</td>
</tr>
<tr>
<td>BEP</td>
<td>.753</td>
<td>.860</td>
<td>.820</td>
<td>.878</td>
<td>.920</td>
</tr>
<tr>
<td>decision rules</td>
<td>ada-boost</td>
<td>ANN</td>
<td>ada-boost</td>
<td>SVM</td>
<td></td>
</tr>
</tbody>
</table>

- … better than people
Sentiment Categorization

• Background and motivation

Sentiment categorization
  – Text categorization

• Applied standard text categorization algorithms
  – Features = bag of words
  – Classifier = machine learning algorithms
    • SVMs, naïve Bayes, maxent

• Data
  – Reviews from IMDb archive
  – 752 negative, 1301 positive

Sentiment categorization appears to be harder than categorizing by topic, e.g. ~82% accuracy for movie reviews.

[E.g., Pang, Lee, Vaithyanathan, EMNLP 2002]

Human Labeling

<table>
<thead>
<tr>
<th>Proposed word lists</th>
<th>Accuracy</th>
<th>Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 1 positive: dazzling, brilliant, phenomenal, excellent, fantastic</td>
<td>58%</td>
<td>78%</td>
</tr>
<tr>
<td>negative: suck, terrible, awful, unwatchable, horrid</td>
<td>64%</td>
<td>39%</td>
</tr>
<tr>
<td>Human 2 positive: gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting negative: bad, dud, sucks, boring, stupid, slow</td>
<td>58%</td>
<td>78%</td>
</tr>
</tbody>
</table>

ML Results

<table>
<thead>
<tr>
<th>Features</th>
<th># of features</th>
<th>frequency or presence?</th>
<th>NB</th>
<th>ME</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) unigrams</td>
<td>16165</td>
<td>freq</td>
<td>78.7</td>
<td>N/A</td>
<td>72.8</td>
</tr>
<tr>
<td>(2) unigrams</td>
<td>5</td>
<td>pres.</td>
<td>81.0</td>
<td>80.4</td>
<td>82.9</td>
</tr>
<tr>
<td>(3) unigrams+bigrams</td>
<td>32330</td>
<td>pres.</td>
<td>80.6</td>
<td>80.8</td>
<td>82.7</td>
</tr>
<tr>
<td>(4) bigrams</td>
<td>16165</td>
<td>pres.</td>
<td>77.3</td>
<td>77.4</td>
<td>77.1</td>
</tr>
<tr>
<td>(5) unigrams+POS</td>
<td>16695</td>
<td>pres.</td>
<td>81.5</td>
<td>80.4</td>
<td>81.9</td>
</tr>
<tr>
<td>(6) adjectives</td>
<td>2633</td>
<td>pres.</td>
<td>77.0</td>
<td>77.7</td>
<td>75.1</td>
</tr>
<tr>
<td>(7) top 2633 unigrams</td>
<td>2633</td>
<td>pres.</td>
<td>80.3</td>
<td>81.0</td>
<td>81.4</td>
</tr>
<tr>
<td>(8) unigrams+position</td>
<td>22430</td>
<td>pres.</td>
<td>81.0</td>
<td>80.1</td>
<td>81.6</td>
</tr>
</tbody>
</table>

3-fold cross-validation
Baselines: 50% - 69%

Figure 1: Baseline results for human word lists. Data: 700 positive and 700 negative reviews.
What is the problem?

– This laptop is *a great deal*.
– *A great deal* of media attention surrounded the release of the new laptop model.
– If you think this laptop is *a great deal*, I’ve got a nice bridge for you to buy.

– …The protagonist tries to protect her *good* name...

[Examples from Lillian Lee]