Text Data Mining:
Predictive and Exploratory Analysis of Text

Jaime Arguello
jarguell@email.unc.edu
Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications
Introductions

- Hello, my name is ______.
- I’m in the ______ program.
- I’m taking this course because I’d like to learn how to ______.
What is Text Data Mining?

- The science and practice of building and evaluating computer programs that automatically detect or discover interesting and useful things in collections of natural language text.
Related Fields

• Machine Learning: developing computer programs that improve their performance with “experience”
• Data Mining: developing methods that discover patterns within large structured datasets
• Statistics: developing methods for the interpretation of data and experimental outcomes in reaching conclusions with a certain degree of confidence
Text Data Mining in this Course

- Predictive Analysis of Text
  - developing computer programs that automatically recognize or detect a particular concept within a span of text

- Exploratory Analysis of Text:
  - developing computer programs that automatically discover interesting and useful patterns or trends in text collections
Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications
Predictive Analysis
example: recognizing triangles
We could imagine writing a “triangle detector” by hand:

- if shape has three sides, then shape = triangle.
- otherwise, shape = other

Alternatively, we could use supervised machine learning!
Predictive Analysis
example: recognizing triangles

training

labeled examples

machine learning algorithm

model

predictions

testing

new, unlabeled examples
Predictive Analysis

example: recognizing triangles

training

What is the part that is missing?

labeled examples

new, unlabeled examples

model

predictions
Predictive Analysis
raw data
### Predictive Analysis
representation: features

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th># sides</th>
<th>equal sides</th>
<th>...</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>yes</td>
</tr>
</tbody>
</table>
Predictive Analysis
example: recognizing triangles

**labeled examples**

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>sides</th>
<th>equal sides</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

**training**

machine learning algorithm

**model**

**testing**

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>sides</th>
<th>equal sides</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td>???</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>???</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td>???</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td>???</td>
</tr>
</tbody>
</table>

**predictions**

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>sides</th>
<th>equal sides</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Predictive Analysis
basic ingredients

1. Training data: a set of examples of the concept we want to automatically recognize

2. Representation: a set of features that we believe are useful in recognizing the desired concept

3. Learning algorithm: a computer program that uses the training data to learn a predictive model of the concept
Predictive Analysis

basic ingredients

1. **Training data**: examples of the concept we want to automatically recognize

2. **Representation**: a set of features that we believe are useful in recognizing the desired concept

3. **Learning algorithm**: a computer program that uses the training data to learn a predictive model of the concept

Highly influential!
Predictive Analysis
basic ingredients

4. **Model**: a (mathematical) function that describes a predictive relationship between the feature values and the presence/absence of the concept

5. **Test data**: a set of previously unseen examples used to estimate the model’s effectiveness

6. **Performance metrics**: a set of statistics used to measure the predictive effectiveness of the model
Predictive Analysis
basic ingredients: the focus in this course

1. **Training data:** a set of examples of the concept we want to automatically recognize

2. **Representation:** a set of features that we believe are useful in recognizing the desired concept

3. **Learning algorithm:** uses the training data to learn a predictive model of the “concept”
Predictive Analysis
basic ingredients: the focus in this course

4. Model: describes a predictive relationship between feature values and the presence/absence of the concept

5. Test data: a set of previously unseen examples used to estimate the model’s effectiveness

6. Performance metrics: a set of statistics used to measure the predictive effectiveness of the model
Predictive Analysis
applications

• Topic categorization
• Opinion mining
• Sentiment analysis
• Bias or viewpoint detection
• Discourse analysis
• Forecasting and nowcasting
• Any other ideas?
Predictive Analysis
example: recognizing triangles

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>sides</th>
<th>equal sides</th>
<th>...</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>yes</td>
</tr>
</tbody>
</table>

labeled examples

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>sides</th>
<th>equal sides</th>
<th>...</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
</tbody>
</table>

new, unlabeled examples

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>sides</th>
<th>equal sides</th>
<th>...</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>???</td>
</tr>
</tbody>
</table>

machine learning algorithm

model

predictions
What Could Possibly Go Wrong?

1. Bad feature representation
2. Bad data + misleading correlations
3. Noisy labels for training and testing
4. Bad learning algorithm
5. Misleading evaluation metric
Training data + Representation
what could possibly go wrong?
## Training data + Representation

what could possibly go wrong?

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>90 deg. angle</th>
<th>equal sides</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Training data + Representation
what could possibly go wrong?

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th>90 deg. angle</th>
<th>equal sides</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>big</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>green</td>
<td>big</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>blue</td>
<td>small</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>red</td>
<td>big</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

1. bad feature representation!
Training data + Representation
what could possibly go wrong?
## Training data + Representation

what could possibly go wrong?

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th># sides</th>
<th>equal sides</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>red</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>green</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Training data + Representation
what could possibly go wrong?

2. bad data + misleading correlations

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th># sides</th>
<th>equal sides</th>
<th>...</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>red</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>green</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td></td>
<td>no</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>yes</td>
</tr>
</tbody>
</table>

2. bad data + misleading correlations
Training data + Representation
what could possibly go wrong?
### Training data + Representation

**what could possibly go wrong?**

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th># sides</th>
<th>equal sides</th>
<th>...</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>white</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td>...</td>
<td>yes</td>
</tr>
<tr>
<td>white</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td>...</td>
<td>no</td>
</tr>
<tr>
<td>white</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td>...</td>
<td>yes</td>
</tr>
<tr>
<td>white</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td>...</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td>...</td>
<td>yes</td>
</tr>
</tbody>
</table>

#### 3. noisy training data!
Learning Algorithm + Model

what could possibly go wrong?

- Linear classifier

\[ y = \begin{cases} 
1 & \text{if } w_0 + \sum_{j=1}^{n} w_j x_j > 0 \\
0 & \text{otherwise} 
\end{cases} \]
Learning Algorithm + Model
what could possibly go wrong?

- Linear classifier

\[
y = \begin{cases} 
1 & \text{if } w_0 + \sum_{j=1}^{n} w_j x_j > 0 \\
0 & \text{otherwise}
\end{cases}
\]

parameters learned by the model
predicted value (e.g., 1 = positive, 0 = negative)
Learning Algorithm + Model
what could possibly go wrong?

test instance

<table>
<thead>
<tr>
<th>f_1</th>
<th>f_2</th>
<th>f_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

model parameters

<table>
<thead>
<tr>
<th>w_0</th>
<th>w_1</th>
<th>w_2</th>
<th>w_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-5</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

output = 2.0 + (0.50 \times -5) + (1 \times 2) + (0.2 \times 1)

output = 1.7

output prediction = positive
Learning Algorithm + Model
what could possibly go wrong?

Learning Algorithm + Model
what could possibly go wrong?

- Would a linear classifier do well on positive (black) and negative (white) data that looks like this?
Learning Algorithm + Model
what could possibly go wrong?

4. Bad learning algorithm!
Learning Algorithm + Model
what could possibly go wrong?

\[ X_1 > 0.50 \]

\[ \begin{align*}
    &\text{yes} &\text{no} \\
    &X_2 > 0.5 &X_2 > 0.5 \\
    &\text{yes} &\text{no} \\
    &\text{pos} &\text{neg} &\text{neg} &\text{pos}
\end{align*} \]
Most evaluation metrics can be understood using a contingency table.

<table>
<thead>
<tr>
<th>True</th>
<th>Triangle</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangle</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Other</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

- What number(s) do we want to maximize?
- What number(s) do we want to minimize?
Evaluation Metric
what could possibly go wrong?

- **True positives (A):** number of triangles correctly predicted as triangles
- **False positives (B):** number of “other” incorrectly predicted as triangles
- **False negatives (C):** number of triangles incorrectly predicted as “other”
- **True negatives (D):** number of “other” correctly predicted as “other”
Evaluation Metric
what could possibly go wrong?

- **Accuracy**: percentage of predictions that are correct (i.e., true positives and true negatives)

\[
\frac{(\text{?} + \text{?})}{(\text{?} + \text{?} + \text{?} + \text{?} + \text{?})}
\]

<table>
<thead>
<tr>
<th>predicted</th>
<th>true</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>triangle</td>
<td>triangle</td>
<td>A</td>
</tr>
<tr>
<td>triangle</td>
<td>other</td>
<td>B</td>
</tr>
<tr>
<td>other</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>
Evaluation Metric
what could possibly go wrong?

- **Accuracy**: percentage of predictions that are correct (i.e., true positives and true negatives)

\[
\frac{(A + D)}{(A + B + C + D)}
\]

<table>
<thead>
<tr>
<th></th>
<th>true</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>predicted</td>
<td>triangle</td>
<td>other</td>
</tr>
<tr>
<td>triangle</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>other</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>
• **Accuracy:** percentage of predictions that are correct (i.e., true positives and true negatives)

• What is the accuracy of this model?
Interpreting the value of a metric on a particular data set requires some thinking ...

On this dataset, what would be the expected accuracy of a model that does NO learning (degenerate baseline)?
Interpreting the value of a metric on a particular data set requires some thinking ...

5. Misleading interpretation of a metric value!
What Could Possibly Go Wrong?

1. Bad feature representation
2. Bad data + misleading correlations
3. Noisy labels for training and testing
4. Bad learning algorithm
5. Misleading evaluation metric
Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications
Text Data Mining in this Course

- Predictive Analysis of Text
  - developing computer programs that automatically recognize a particular concept within a span of text

- Exploratory Analysis of Text:
  - developing computer programs that automatically discover useful patterns or trends in text collections
Exploratory Analysis

example: clustering shapes
Exploratory Analysis
example: clustering shapes
Exploratory Analysis
example: clustering shapes

unlabeled examples

clustering algorithm
# Exploratory Analysis
representation: features

<table>
<thead>
<tr>
<th>color</th>
<th>size</th>
<th># sides</th>
<th>equal sides</th>
<th>...</th>
<th>shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>no</td>
<td></td>
<td>triangle</td>
</tr>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>triangle</td>
</tr>
<tr>
<td>red</td>
<td>small</td>
<td>inf</td>
<td>yes</td>
<td></td>
<td>circle</td>
</tr>
<tr>
<td>green</td>
<td>small</td>
<td>4</td>
<td>yes</td>
<td></td>
<td>square</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>blue</td>
<td>big</td>
<td>3</td>
<td>yes</td>
<td></td>
<td>triangle</td>
</tr>
</tbody>
</table>
Exploratory Analysis
basic ingredients

1. **Data:** a set of examples that we want to automatically analyze in order to discover interesting trends

2. **Representation:** a set of features that we believe are useful in describing the data (i.e., its main attributes)

3. **Similarity Metric:** a measure of similarity between two examples that is based on their feature values

4. **Clustering algorithm:** an algorithm that assigns items with similar feature values to the same group
Representation
what could possibly go wrong?
Representation
what could possibly go wrong?
Exploratory Analysis

basic ingredients: the focus in this course

1. **Data**: a set of examples that we want to automatically analyze in order to discover interesting trends

2. **Representation**: a set of features that we believe are useful in describing the data

3. **Similarity Metric**: a measure of similarity between two examples that is based on their feature values

4. **Clustering algorithm**: an algorithm that assigns items with similar feature values to the same group
Text Data Mining in this Course

• Predictive Analysis of Text
  ‣ developing computer programs that automatically recognize or detect a particular concept within a span of text

• Exploratory Analysis of Text:
  ‣ developing computer programs that automatically discover interesting and useful patterns or trends in text collections
Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications
Predictive Analysis of Text
examples we’ll cover in class

• Topic Categorization
• Opinion Mining
• Sentiment/Affect Analysis
• Bias Detection
• Information Extraction and Relation Learning
• Text-driven Forecasting
• Temporal Summarization
Predictive Analysis of Text
example applications

- **Topic Categorization**: automatically assigning documents to a set of pre-defined topical categories
Topic Categorization
Topic Categorization

Arts
- Movies, Television, Music...

Games
- Video Games, RPGs, Gambling...

Kids and Teens
- Arts, School Time, Teen Life...

Reference
- Maps, Education, Libraries...

Shopping
- Clothing, Food, Gifts...

World
- Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web

Copyright © 2012 Netscape
Opinion Mining: automatically detecting whether a span of opinionated text expresses a positive or negative opinion about the item being judged
Opinion Mining
movie reviews

- “Great movie! It kept me on the edge of my seat the whole time. I IMAX-ed it and have no regrets.”
- “Waste of time! It sucked!”
- “This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.”
- “Trust me, this movie is a masterpiece .... after you’ve seen it 4+ times.”
Predictive Analysis of Text

example applications

• **Sentiment/Affect Analysis**: automatically detecting the emotional state of the author of a span of text (usually from a set of pre-defined emotional states).
• “[I] also found out that the radiologist is doing the biopsy, not a breast surgeon. I am more scared now than when I ...”

• “... My radiologist ‘assured’ me my scan was NOT going to be cancer...she was wrong.”

• “ ... My radiologist did my core biopsy. Not a problem and he did a super job of it.”

• “It's pretty standard for the radiologist to do the biopsy so I wouldn't be concerned on that score.”

Sentiment Analysis
support group posts

fear
despair

hope

hope
• **Bias detection**: automatically detecting whether the author of a span of text favors a particular viewpoint (usually from a set of pre-defined viewpoints)
Bias Detection

• “Coming [up] next, drug addicted pregnant women no longer have anything to fear from the authorities thanks to the Supreme Court. Both sides on this in a moment.” -- Bill O’Reilly

• “Nationalizing businesses, nationalizing banks, is not a solution for the democratic party, it's the objective.” -- Rush Limbaugh

• “If you're keeping score at home, so far our war in Iraq has created a police state in that country and socialism in Spain. So, no democracies yet, but we're really getting close.” -- Jon Stewart
Predictive Analysis of Text
example applications

- **Information extraction**: automatically detecting that a short sequence of words belongs to (or is an instance of) a particular entity type, for example:
  - Person(X)
  - Location(X)
  - TennisPlayer(X)
  - ...
• **Relation Learning:** automatically detecting pairs of entities that share a particular relation, for example:

  ‣ CEO(<person>,<company>)
  ‣ Capital(<city>,<country>)
  ‣ Mother(<person>,<person>)
  ‣ ConvictedFelon(<person>,<crime>)
  ‣ ...

Predictive Analysis of Text

example applications
Relation Learning

CEO(<person>, <company>)

<person>, who was named CEO of <company>
Relation Learning

CEO(<person>,<company>)

CEO(Tom LaSorda, Fisker)

CEO(Sean Connolly, Hillshire Brands)

CEO(woman, Gilt Groupe)

CEO(scottish chemist, AztraZeneca)

CEO(Bob Harrison, First Hawaiian Bank)
Text-based Forecasting: monitoring incoming text (e.g., tweets) and making predictions about external, real-world events or trends, for example:

- a presidential candidate’s poll rating
- a company’s stock value change
- a movie’s box office earnings
- side-effects for a particular drug
- ...
Predictive Analysis of Text

example applications

• **Temporal Summarization:** monitoring incoming text (e.g., tweets) about a news event and predicting whether a sentence should be included in an on-going summary of the event

• Updates to the summary should contain relevant, novel, and accurate information.

(output)

\[ S_{t-1} \]

(input)

\[ B_I \]

(Diagram from Guo et al., ECIR 2013)
Detecting other interesting properties of text: [insert your crazy idea here], for example, detecting humorous text:

- “Beauty is in the eye of the beholder” not funny
- “Beauty is in the eye of the beer holder” funny

(example from Mihalcea and Pulman, 2007)
Outline

Introductions

What is Text Data Mining?

Predictive Analysis of Text: The Big Picture

Exploratory Analysis of Text: The Big Picture

Applications
Course Overview
Road Map
first half of the semester

- Predictive Analysis of Text
  - Supervised machine learning principles
  - Text representation
  - Feature selection
  - Basic machine learning algorithms
  - Tools for predictive analysis of text
  - Experimentation and evaluation

- Exploratory Analysis of Text
  - Clustering
  - Outlier detection (tentative)
  - Co-occurrence statistics
Road Map
second half of the semester

- Applications
  - Text classification
  - Opinion mining
  - Sentiment analysis
  - Bias detection
  - Information extraction
  - Relation learning
  - Text-based forecasting
  - Temporal Summarization

- Is there anything that you would like to learn more about?
Grading

- 30% homework
  - 10% each
- 20% midterm
- 40% term project
  - 5% proposal
  - 10% presentation
  - 25% paper
- 10% participation
Grading for Graduate Students

• H: 95-100%
• P: 80-94%
• L: 60-79%
• F: 0-59%
Grading for Undergraduate Students

- A+: 97-100%
- A: 94-96%
- A-: 90-93%
- B+: 87-89%
- B: 84-86%
- B-: 80-83%
- C+: 77-79%
- C: 74-76%
- C-: 70-73%
- D+: 67-69%
- D: 64-66%
- D-: 60-63%
- F: <= 59%
General Outline of Homework

• Given a dataset (i.e., a training and test set), run experiments where you try to predict the target class using different feature representations
• Do error analysis
• Report on what worked, what didn’t, and why!
• Answer essay questions about the assignment
  ‣ These will be associated with the course material
Homework vs. Midterm

- The homework will be more challenging than the midterm. It should be, you have more time.
Course Tips

- Work hard
- Do the assigned readings
- Do other readings
- Be patient and have reasonable expectations
  - you’re not supposed to understand everything we cover in class during class
- Seek help sooner rather than later
  - office hours: by appointment
  - questions via email
- Remember the golden rule: no pain, no gain
Questions?