CS544: Sentiments, Reviews and Regression Analysis

March 29, 2012

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

What is subjectivity?

• The linguistic expression of somebody’s opinions, sentiments, emotions, evaluations, beliefs, speculations (private states)

• Private state: state that is not open to objective observation or verification

• Subjectivity analysis classifies content in objective or subjective

Search Engines

The Best Tablet: Kindle Fire or iPad 2?
www.squidoo.com ... > Gadgets > Unique Gadgets
Kindle Fire, the new tablet developed by Amazon is here, and it is currently the … Ipad is way better, and now that they have come out with an Ipad 3 the Fire is …
Today

- Product review mining:
  What are the features of the iPad3 that the customers like?
- Review classification:
  Is a review positive or negative toward the movie?
- Tracking sentiments toward topics over time:
  Is my wife still angry?
- Prediction (election outcomes, market trends):
  Will Clinton or Obama win?
- Text summarization

What is sentiment analysis?

- Also known as opinion mining
- Attempts to identify the opinion/sentiment that a person may hold towards an object
- It is a finer grain analysis compared to subjectivity analysis

<table>
<thead>
<tr>
<th>Sentiment Analysis</th>
<th>Subjectivity analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Subjective</td>
</tr>
<tr>
<td>Negative</td>
<td>Objective</td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
</tr>
</tbody>
</table>

Example

- A review:
  (1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. (7) She also thought the phone was too expensive, and wanted me to return it to the shop...
• **Object**: an entity that can be a product, service, individual, organization, event, or topic.
  – Ex. iPhone

• **Attribute**: an object usually has two types of attributes
  – components: battery, keypad/touch screen
  – properties: size, weight, color, voice quality

• **Explicit and implicit attributes**:  
  – explicit appear in the opinion:  
    “the battery life of this phone is too short”  
  – implicit do not appear in the opinion:  
    “this phone is too large” (on attribute size)

• **Opinion holder**: the person or organization that expresses the opinion

• **Opinion orientation (polarity)**: positive, negative, or neutral

• **Opinion strength**: level/scale/intensity of opinion indicating how strong it is
  Ex. contented->happy->joyous->ecstatic

---

**Simple Method**

• **Opinion/polarity words**: dominating indicators of sentiments, especially adjectives, adverbs, and verbs, e.g. “I absolutely love this camera. It is amazing!”

• **Pre-defined opinion words**: good, terrible

• **Assign orientation score (+1, -1) to all words**:
  – positive opinion words (+1): great, amazing, love
  – negative opinion words (-1): horrible, hate
  – strength value [0, 1] can be used too

• The orientation score of paragraph/document is the sum of orientation scores of all opinion words found
Available Resources

- **Lexicons**
  - General Inquirer (Stone et al., 1966)
  - OpinionFinder lexicon (Wiebe & Riloff, 2005)
  - SentWordNet (Esuli & Sebastiani, 2006)

- **Annotated corpora**
  - Used in statistical approaches (Hu & Liu 2004, Pang & Lee 2004)
  - MPQA corpus (Wiebe et al., 2005)

- **Tools**
  - Algorithm based on minimum cuts (Pang & Lee, 2004)
  - OpinionFinder (Wiebe et al., 2005)

---

WordNetAffect

- Semi-automatic Method (Stapparava and colleagues)
  - assigned affect labels to words in WordNet
  - expanded the lists using WordNet relations such as synonymy, antonymy, entailment, hyponymy

- Available:
  - [http://wndomains.itc.it/](http://wndomains.itc.it/)

---

<table>
<thead>
<tr>
<th>Affect</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>emotions</td>
<td>noun anger#1, verb fear#1</td>
</tr>
<tr>
<td>cognitive state</td>
<td>noun confusion#2, adjective dazed#2</td>
</tr>
<tr>
<td>physical state</td>
<td>noun illness#1, adjective all#1</td>
</tr>
<tr>
<td>hedonic signal</td>
<td>noun hurt#1, noun suffering#1</td>
</tr>
<tr>
<td>attitude</td>
<td>noun intolerance#1, noun defensive#1</td>
</tr>
<tr>
<td>sensation</td>
<td>noun coldness#1, verb fear#3</td>
</tr>
</tbody>
</table>
How did they Build it?

- Assume that contexts are coherent, i.e. words of the same orientation like to co-occur together, then the presence of one makes the other more probable

- Assume that alternatives are similarly subjective

WordNet

- (72) (adjective) **brilliant, smart as a whip** (having or marked by unusual and impressive intelligence) "some men dislike brusque women"; "a brilliant mind"; "a brilliant solution to the problem"
  - similar to
  - (73) (adjective) **brilliant** (having the capacity for thought and reason especially at a high degree) "These intelligent children"; "an intelligent question"
  - etymology
  - "(adj) (adjective) [related to: **intelligence** (a quality that enables the mind)"
  - (80) (adjective) **intelligent** (relating to intelligence) "a dull child with lazy and unintelligent co-workers"
WordNet

- (7) (appropriate) brilliant, excel as a clerk: (having or marked by unusual and impressive intelligence) "some people dislike brainy men"; "a brilliant mind"; "a brilliant solution to the problem"
- entailment
  - (4) intelligence (having the capacity for thought and reason especially to a high degree) "a brilliant man"; "a brilliant solution to the problem"
- demarcated boundaries
- to (brilliant) (a quality that combines the usual)
- to (brilliance) (ornamental mental ability)
- antonym
  - (4) unbrilliant (a quality that combines the usual)
  - to (brilliance) (ornamental mental ability)

WordNet

- (7) (appropriate) brilliant, excel as a clerk: (having or marked by unusual and impressive intelligence) "some people dislike brainy men"; "a brilliant mind"; "a brilliant solution to the problem"
- entailment
  - (4) intelligence (having the capacity for thought and reason especially to a high degree) "a brilliant man"; "a brilliant solution to the problem"
- demarcated boundaries
- to (brilliant) (a quality that combines the usual)
- to (brilliance) (ornamental mental ability)
- antonym
  - (4) unbrilliant (a quality that combines the usual)
  - to (brilliance) (ornamental mental ability)
WordNet Examples

- A (adj): bright, brilliant, smart, alert, shrewd, acute, keen, alert as a fox, seeing or marked by mental and imaginative agility: "some men with mental acuity and agility"
- bright: [adj]: having or marked by mental and imaginative agility: "some men with mental acuity and agility"
- brilliant: [adj]: seeing or marked by mental and imaginative agility: "some men with mental acuity and agility"

SentiWordNet

- Method (Esuli and Sebastiani):
  - Trained classifiers to give positive, negative, objective rating for each synset in WordNet 2.0
  - The score of a rating is from 0 – 1
- Available:
  - http://sentiwordnet.isti.cnr.it

SentiWordNet
Automatically Finding Opinion Words

- **Data**: reviews from opinions.com on automobiles, banks, moves, and travel
- **Step 1**: run part of speech (POS) tagger and extract two consecutive words if the first word is adjective/adverb and the second one is a noun
  - “this camera produces beautiful pictures”
- **Step 2**: estimate the orientation of each extracted phrase using PMI

\[
\text{sent} = \text{PMI}(\text{adj}, \text{adv}) - \text{PMI}(\text{adv}, \text{adj})
\]

\[
\text{sent} = \log \frac{\text{PMI}(\text{adj}, \text{adv})}{\text{PMI}(\text{adv}, \text{adj})}
\]

Example of Produced Extractions

- “low fees”, “JJ NNS”, 0.333
- “unethical practices”, “JJ NNS”, -0.484
- “low funds”, “JJ NNS”, -0.843
Reviews

- Numerous reviews of varying quality exist
  - how to rank them by importance (helpfulness)
  - how to predict the helpfulness of a newly written review

- Most reviews are ranked by
  - number of stars in Amazon
  - humans

Task Definition

- Given a set of reviews $R$ for a particular product, can we rank the reviews according to their helpfulness

$$h(r \in R) = \frac{\text{rating}_r}{\text{rating}_r + \text{rating}_\neg r}$$

[+\text{rating}_r and \text{-rating}_\neg r indicate number of people finding the review helpful/unhelpful]

Francis Galton (1822–1911)

- Interested in studying the qualification of intellect and disposition which lead to reputation
- Did not have data on “intelligence”, instead studied height correlation between parents and adult children

Cyril Burt (1883 – 1971)

- Interested in studying the IQs of identical twins raised in “foster” homes from IQs of siblings raised in biological parent’s homes
Regression Analysis

- Regression analysis is used to predict the values of one variable (the dependent variable) on the basis of other variables (the independent variables)
  - dependent variable is what we wish to explain
  - independent variable is used to explain the dependent one
- If we have only ONE independent variable, the model
  \[ y = \beta_0 + \beta_1 x + \varepsilon \]
is referred to as simple linear regression

Simple Linear Regression Example

- Let \( y \) be a student’s college achievement, measured by his/her GPA. This might be a function of several variables:
  - \( x_1 \) = rank in class
  - \( x_2 \) = high school overall rating
  - \( x_3 \) = high school GPA
  - \( x_4 \) = SAT scores
  - ...
- We want to predict \( y \) using the knowledge of \( x_1, x_2, x_3 \) and \( x_4 \)
Population Linear Regression

The population regression model:

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

Linear Regression Assumptions

- Error values (\( \varepsilon \)) are statistically independent
- Error values are normally distributed for any given value of \( x \)
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the \( x \) variable and the \( y \) variable is linear

Population Linear Regression

![Diagram of population linear regression model with observed and predicted values, slope, and random error for a given \( x \) value.]}
Which line has the best “fit” to the data?

Setting up the minimization problem

- We want to minimize the sum of the mean squared error in the equation:
  \[ Y_i = b_0 + b_1 X_{i1} + e_i \text{ for observations } i = 1, ..., n \]

- If we rearrange terms:
  \[ e_i = Y_i - b_0 - b_1 X_{i1} \text{ and } e_i^2 = (Y_i - b_0 - b_1 X_{i1})^2 \]
Setting up the minimization problem

- We want to minimize the sum of the mean squared error in the equation:
  \[ Y_i = b_0 + b_1 X_{ij} + e_i \]
  for observations \( i = 1, \ldots, n \)

- If we rearrange terms:
  \[ e_i = Y_i - b_0 - b_1 X_{ij} \]
  and \( e_i^2 = (Y_i - b_0 - b_1 X_{ij})^2 \)

- To minimize the sum of the squared error, we need to minimize the following equation:
  \[ \sum e_i^2 = \sum (Y_i - b_0 - b_1 X_{ij})^2 \]
  with respect to the regression coefficients \( b_0 \) and \( b_1 \).

The Least Squares Equation

- The formulas for \( b_1 \) and \( b_0 \) are:
  \[
  b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}
  \]
  algebraic equivalent:
  \[
  b_1 = \frac{\sum xy - \frac{1}{n} \sum x \sum y}{\sum x^2 - \frac{1}{n} (\sum x)^2}
  \]
  and
  \[
  b_0 = \bar{y} - b_1 \bar{x}
  \]

Interpretation

- \( b_0 \) (intercept) is the estimated average value of \( y \) when the value of \( x \) is zero

- \( b_1 \) (slope) is the estimated change in the average value of \( y \) as a result of a one-unit change in \( x \)
Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- Let's select a random sample of 10 houses
  - Dependendent variable ($y$) = house price in $1000s
  - Independent variable ($x$) = square feet

Sample Data for House Price Model

<table>
<thead>
<tr>
<th>House Price In $1000s ($Y$)</th>
<th>Square Feet ($X$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>1400</td>
</tr>
<tr>
<td>312</td>
<td>1600</td>
</tr>
<tr>
<td>279</td>
<td>1700</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
</tr>
<tr>
<td>199</td>
<td>1100</td>
</tr>
<tr>
<td>219</td>
<td>1550</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
</tr>
<tr>
<td>319</td>
<td>1425</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
</tr>
</tbody>
</table>

Graphical Representation of House Price Model
### Computation

\[ \hat{b}_1 = \frac{\sum xy - \sum x \sum y}{\sum x^2 - \left(\sum x\right)^2/n} \]

\[ \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( y^2 )</th>
<th>( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>1400</td>
<td>60025</td>
<td>1960000</td>
</tr>
<tr>
<td>312</td>
<td>1600</td>
<td>97434</td>
<td>2560000</td>
</tr>
<tr>
<td>279</td>
<td>1700</td>
<td>77841</td>
<td>2890000</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
<td>94684</td>
<td>3515625</td>
</tr>
<tr>
<td>199</td>
<td>1100</td>
<td>39601</td>
<td>1210000</td>
</tr>
<tr>
<td>219</td>
<td>1550</td>
<td>47961</td>
<td>2402500</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
<td>144025</td>
<td>5522500</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
<td>104976</td>
<td>6002500</td>
</tr>
<tr>
<td>319</td>
<td>1425</td>
<td>101761</td>
<td>2030625</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
<td>65025</td>
<td>2890000</td>
</tr>
</tbody>
</table>

**SUM**

- \( \sum y = 2865 \)
- \( \sum x = 17150 \)

- \( \sum y^2 = 853423 \)
- \( \sum x^2 = 30983750 \)
Computation

\[ b_0 = \bar{y} - b_1 \bar{x} \]

\[ b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \]

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>Y^2</th>
<th>X^2</th>
<th>X*Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>140</td>
<td>60025</td>
<td>196000</td>
<td>34300</td>
</tr>
<tr>
<td>312</td>
<td>160</td>
<td>97344</td>
<td>256000</td>
<td>49920</td>
</tr>
<tr>
<td>279</td>
<td>170</td>
<td>77841</td>
<td>289000</td>
<td>47430</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
<td>94864</td>
<td>3515625</td>
<td>577500</td>
</tr>
<tr>
<td>199</td>
<td>1100</td>
<td>38601</td>
<td>1210000</td>
<td>218900</td>
</tr>
<tr>
<td>219</td>
<td>1550</td>
<td>47861</td>
<td>2402500</td>
<td>339450</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
<td>164025</td>
<td>5522500</td>
<td>951750</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
<td>104976</td>
<td>6002500</td>
<td>793800</td>
</tr>
<tr>
<td>319</td>
<td>1425</td>
<td>191761</td>
<td>2030625</td>
<td>454575</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
<td>65025</td>
<td>2890000</td>
<td>433500</td>
</tr>
<tr>
<td>2865</td>
<td>17150</td>
<td>563423</td>
<td>10883750</td>
<td>5089975</td>
</tr>
</tbody>
</table>

\( b_0 = 98.248 \)  
\( b_1 = 0.1097 \)

Graphical Representation of House Price Model

\[ \text{Intercept} = 98.248 \]  
\[ \text{Slope} = 0.10977 \]

\[ \text{house price} = 98.248 + 0.10977 \text{ (square feet)} \]

Interpretation of the Intercept

\[ \text{house price} = 98.248 + 0.10977 \text{ (square feet)} \]

- \( b_0 \) is the estimated average value of \( Y \) when the value of \( X \) is zero (if \( x = 0 \) is in the range of observed \( x \) values)

In the example no houses had 0 square feet, so \( b_0 = 98.24833 \) just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet
Interpretation of the Slope Coefficient

$$\text{house price} = 98.248 + 0.10977 \text{ (square feet)}$$

- \(b_1\) measures the estimated change in the average value of \(Y\) as a result of a one-unit change in \(X\)

\(b_1 = .10977\) tells us that the average value of a house increases by \(.10977(1000) = 109.77\), on average, for each additional one square foot of size.

Predict the price for a house with 2000 square feet

<table>
<thead>
<tr>
<th>House Price in $1000s</th>
<th>Square Feet (ft)</th>
</tr>
</thead>
<tbody>
<tr>
<td>245</td>
<td>1400</td>
</tr>
<tr>
<td>270</td>
<td>1700</td>
</tr>
<tr>
<td>308</td>
<td>1875</td>
</tr>
<tr>
<td>300</td>
<td>1900</td>
</tr>
<tr>
<td>210</td>
<td>1500</td>
</tr>
<tr>
<td>405</td>
<td>2350</td>
</tr>
<tr>
<td>324</td>
<td>2450</td>
</tr>
<tr>
<td>210</td>
<td>1420</td>
</tr>
<tr>
<td>255</td>
<td>1700</td>
</tr>
</tbody>
</table>

Estimated Regression Equation:

- \(\text{house price} = 98.25 + 0.1098 \text{ (sq.ft.)}\)
- \(\text{house price} = 98.25 + 0.1098(2000) = 317.85\)

This means that the predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850.

Other Applications

- Achievement of a college student based on the time spent studying
- Predict the performance of your NE systems on the test data
  - using the features and the output of your system on the development data set
  - using the development and test results from last year’s students
- Performance of any NLP toolkit (parser accuracy)
- ...