Class 10: Sentiment Analysis

MAST5953: Web Scraping and Text Mining

Dr. Miriam Sorace

www.miriamsorace.net

15 February 2023

Outline of Today's Class

Sentiment Analysis: Definitions & Examples

Sentiment Analysis

What for?

- Detecting the *attitude* of a text
  - positive/negative ...
  - left/right ...
  - anti-EU/pro-EU ...
- A classification exercise, but sentiment analysis can also be continuum/scale

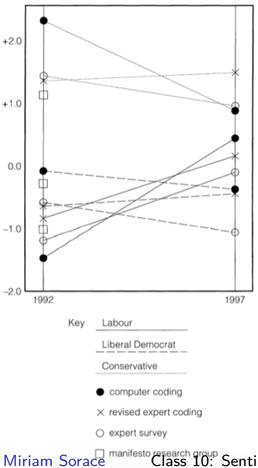
Sentiment Analysis
Methods

1. Dictionaries
   - Generating lists of words for each sentiment category
2. Machine Learning
   - Training an algorithm via pre-labeled documents (e.g. Naive Bayes)

Sentiment Analysis

Example 1: Laver, M., Garry, J. (2000) "Estimating Policy Positions from Political Texts" *American Journal of Political Science*

Figure 1 Standardized Expert Survey, Computer Coded and Expert Coded Estimates of Party Policy Positions in Britain 1992–97



Sentiment Analysis

Example 1: Laver, M., Garry, J. (2000) "Estimating Policy Positions from Political Texts" *American Journal of Political Science*

- **Economy**
- **Institutions**
- **Values**
- **Law and Order**
- **Environment**
- **Culture**
- **Groups**
- **Rural**
- **Urban**

Sentiment Analysis

Example 1: Laver, M., Garry, J. (2000) "Estimating Policy Positions from Political Texts" *American Journal of Political Science*

- **Economy**
  - *+ State*
  - *= State*
  - *- State*
- **Institutions**
- **Values**
- **Law and Order**
- **Environment**
- **Culture**
- **Groups**
- **Rural**
- **Urban**

Sentiment Analysis

Example 1: Laver, M., Garry, J. (2000) "Estimating Policy Positions from Political Texts" *American Journal of Political Science*

- **Economy**
  - $+$ *State*
    - `accommodation; age; ambulance; assist; benefit; care; class; clinics; deprivation; disabilities; disadvantaged; elderly; establish; hardship; hunger; invest; patients; pensions; poor; poverty; school; child; collective; contribution ...`
  - $=$ *State*
  - \- *State*
- **Institutions**
- **Values**
- **Law and Order**
- **Environment**
- **Culture**
- **Groups**
- **Rural**
- **Urban**

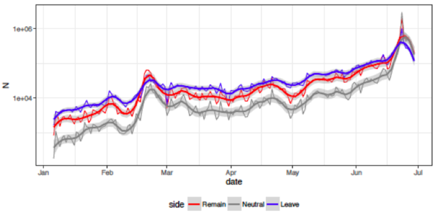
Sentiment Analysis

Example 2: Benoit, K., Matsuo, A. (2018) "Brexit Discussion on Social Media" *The EUEngage Working Paper Series*

Figure 2. Word Cloud of High Frequency Hashtags

Note: Top seventy hashtags in each category. The categories of hash tags are generated from the predicted probability of a hashtag to belong to Leave (0-0.25: Remain, 0.25-0.75: Neutral, 0.75-1: Leave). Size of hashtag is proportional to the frequency. Color indicates the category.

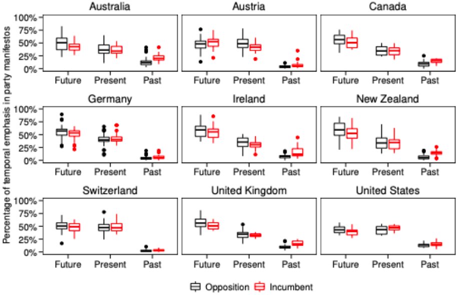
Figure 3 Time-series of Leave and Remain Tweets (Jan June 2016)

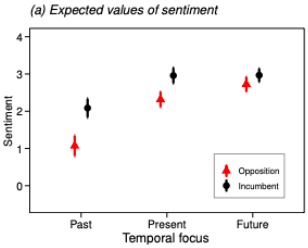


Sentiment Analysis

Example 3: Müller, S (forthcoming)"The Temporal Focus of Campaign Communication" *Journal of Politics*

Figure 1: The emphasis on the past, present, and future, conditional on incumbency status





Sentiment Analysis
Engage with the Examples

1. Explore the Laver/Garry Dictionary yourself at: `http://yoshikoder.sourceforge.net/code/yoshikoder/dictionaries/LaverGarryAJPS.ykd`
2. Watch Ken Benoit's presentation of the Brexit project here `https://www.youtube.com/watch?v=IVayXmtI2VM`

Dictionaries

Dictionary Analysis
How it Works

- Both qualitative and quantitative
  - Contextual knowledge needed: validation crucial
  - Once defined, the dictionary will be automated: perfectly reliable
- Identify key concepts/categories (or "keys")
- Identify words/n-grams (the "values") associated with each key
  - From Laver & Garry:
    - **more state**: `assist, benefit, care, disabilities, educat*,invest, pension`
    - **less state**: `autonomy, bidders, choice*, controls, market`

Dictionary Analysis

How to build one

1. Order your concepts/keys **hierarchically**
   1.1 Domain - *Economy*
   1.2 Sub-Domain - *Labour Law*
   1.3 Sentiment Categories/Poles - *Pro-Business/Neutral/Pro-Worker*
2. Identify extreme texts among the texts with known positions: the "archetypes"
3. Identify words/n-grams that are statistically associated with the various archetypes
   - Chi-square tests
4. Examine these words/n-grams for their specificity: are they polysemes?
5. Examine these words to decide whether stemming is necessary
6. Create word/n-grams lists for the relevant dictionary key
7. Investigate whether the dictionary is sensitive enough: will it capture all instances of [key]?

Dictionary Analysis
Advantages

- Allows for detailed contextual knowledge to be reliably applied to large-scale text analysis

Dictionary Analysis

Disadvantages

- Time-consuming
- Non-generalisable: often dictionaries do not travel well to new corpuses
  - E.g. `freez*` is positive in the context of refrigeration appliances but negative in the context of computing
  - E.g. `revolut*` is positive in the context of technology, negative in the context of interior policy
- Difficult to know with certainty how comprehensive/valid the dictionary is

The Naive Bayes Algorithm

Naive Bayes

How it Works

- Bayes' Rule:

$$P(C_j|W_i) = \frac{P(W_i|C_j)P(C_j)}{P(W_i)}$$

- which can be transformed to:

$$P(C|D) = P(C)\prod\frac{P(W_i|C)}{P(W_i)}$$

Naive Bayes
How it Works: Example

- Training Set:

Document	Words	Class
1	like love fantastic perfect	*Positive*
2	love love great mean	*Positive*
3	awful terrible worse mean	*Negative*
4	like fantastic great like	*Positive*
5	terrible awful love mean	*??*

- What is the likelihood that the new document *5* is of class *Positive* vs. the likelihood that it is of class *Negative*?

Naive Bayes

How it Works: Example

Document	Words	Class
1	like love fantastic perfect	*Positive*
2	love love great mean	*Positive*
3	awful terrible worse mean	*Negative*
4	like fantastic great like	*Positive*
5	terrible awful love mean	*??*

$$P(C_{pos}|D_5) = P(C_{pos})\frac{\prod P(W_{i5}|C_{pos})}{P(W_{i5})}$$

$$= 0.75\frac{(0.04 * 0.04 * 0.29 * 0.13)}{(0.09 + 0.09 + 0.22 + 0.16)}$$

$$= 0.00008$$

Naive Bayes

How it Works: Example

Document	Words	Class
1	like love fantastic perfect	*Positive*
2	love love great mean	*Positive*
3	awful terrible worse mean	*Negative*
4	like fantastic great like	*Positive*
5	terrible awful love mean	*??*

$$P(C_{neg}|D_5) = P(C_{neg})\frac{\prod P(W_{i5}|C_{neg})}{P(W_{i5})}$$

$$= 0.25\frac{(0.38 * 0.38 * 0.13 * 0.38)}{(0.09 + 0.09 + 0.22 + 0.16)}$$

$$= 0.003$$

Naive Bayes
Steps

1. Obtain a valid and reliable labeled set
   - Expert-coded
   - Label from meta-data - e.g. party
   - Crowd-sourced
2. Run the Naive Bayes classifier algorithm
3. Test the performance via cross-validation
   - Accuracy, Recall, Precision, F-Measure

Naive Bayes

Advantages

- Outperforms dictionaries in the sensitivity of classification - as long as the training sample is big enough
- Flexible and quick: it can be easily re-applied to new corpuses, provided satisfactory identification of archetypal training texts

Naive Bayes

Disadvantages

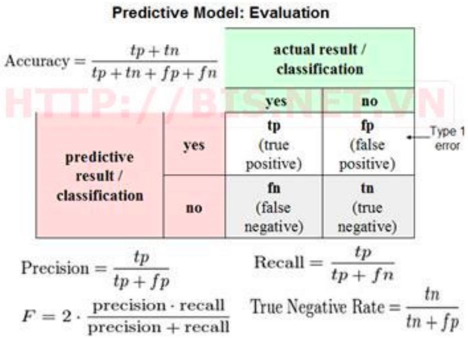
- Naive: word order does not count + probability of words/n-grams assumed independent from the class
  - formula might not correctly model the data-generation process!
- Very reliant on the training set: select a good one!
  - Make sure this is representative of the extreme points
  - Make sure it is large enough so that language styles/rhetoric does not influence classification
  - Make sure to appropriately pre-process the texts!
- Requires a lot of validation ex post (cross-validation steps).

Naive Bayes

Performance Metrics

- **Accuracy**: correctly classified texts divided by the total number of texts
- **Precision**: How many texts that were *predicted* as class A were *actual* class A texts?
- **Recall**: How many texts that *actually* are of class A were also *predicted* to be of class A?
- **F-Measure**: composite measure of precision and recall. Good recall can lead to low precision, so a mix measure is needed.

Naive Bayes
The Confusion Matrix




```
TP : The number of samples of class c are correctly classified into class c
FP: The number of samples not belonging to class c misclassified into class c
TN: The number of samples not belonging to class c is classified (correctly)
FN: The number of samples of class c misclassified (in other classes c)
```

Example from:
https://stats.stackexchange.com/questions/116585/

Sentiment Analysis in R

Sentiment Analysis: R code demonstration

- R Code
- Use your scraped tweets!

Naive Bayes Analysis in R

- For a tutorial check out:
- `https://tutorials.quanteda.io/machine-learning/nb/`