

# Quantitative Text Analysis. Applications to Social Media Research

**Pablo Barberá**

London School of Economics

`www.pablobarbera.com`

Course website:

[pablobarbera.com/text-analysis-vienna](http://pablobarbera.com/text-analysis-vienna)

# Dictionary Methods Applied to Social Media Text

# Dictionary methods

Classifying documents when categories are known:

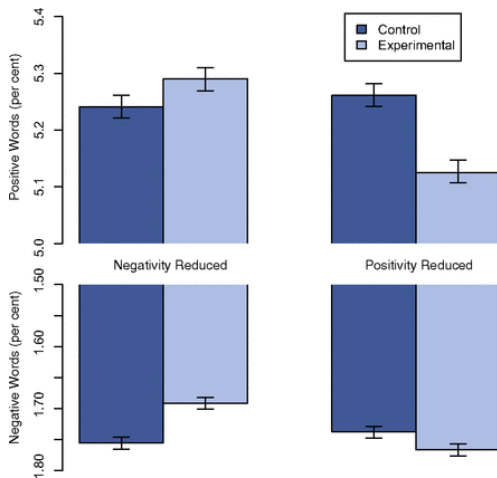
- ▶ Lists of words that correspond to each category:
  - ▶ Positive or negative, for sentiment
  - ▶ Sad, happy, angry, anxious... for emotions
  - ▶ Insight, causation, discrepancy, tentative... for cognitive processes
  - ▶ Sexism, homophobia, xenophobia, racism... for hate speech

many others: see LIWC, VADER, SentiStrength, LexiCoder...
- ▶ Count number of times they appear in each document
- ▶ Normalize by document length (optional)
- ▶ **Validate, validate, validate.**
  - ▶ Check sensitivity of results to exclusion of specific words
  - ▶ Code a few documents manually and see if dictionary prediction aligns with human coding of document

# Linguistic Inquiry and Word Count

- ▶ Created by Pennebaker et al — see <http://www.liwc.net>
- ▶ Uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- ▶ Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- ▶ For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- ▶ Hierarchical: so “anger” are part of an *emotion* category and a *negative emotion* subcategory
- ▶ You can [buy](http://www.liwc.net/descriptiontable1.php) it here:  
<http://www.liwc.net/descriptiontable1.php>

# Example: Emotional Contagion on Facebook



**Source:** Kramer et al, PNAS 2014

# Potential advantage: Multi-lingual

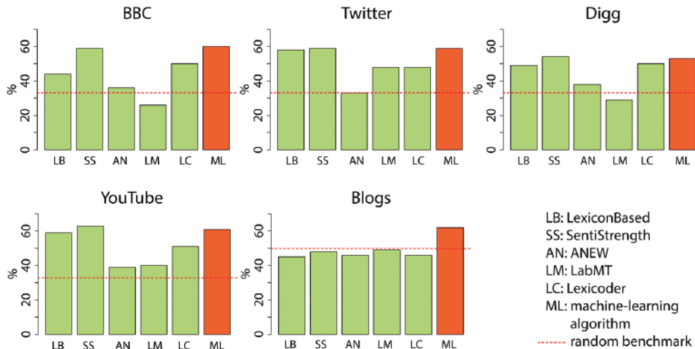
APPENDIX B  
DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

|                | NL  | UK   | GE   | IT   |
|----------------|---|--|--|--|
| <b>Core</b>    | elit*<br>consensus*<br>ondemocratisch*<br>ondemokratisch*<br>referend*<br>corrupt*<br>propagand*<br>politici*<br>*bedrog*<br>*bedrieg*<br><br>*verraa*<br>*verrad*<br>schaam*<br><br>schand*<br>waarheid*<br>oneerlijk* | elit*<br>consensus*<br>undemocratic*<br><br>referend*<br>corrupt*<br>propagand*<br>politici*<br>*deceit*<br>*deceiv*<br><br>*betray*<br><br>shame*<br><br>scandal*<br>truth*<br>dishonest* | elit*<br>konsens*<br>undemokratisch*<br><br>referend*<br>korrupt*<br>propagand*<br>politiker*<br>täusch*<br>betrüg*<br>betrug*<br>*verrat*<br><br>scham*<br>schäm*<br>skandal*<br>wahrheit*<br>unfair*<br>unehrlich*<br>establishm*<br>*herrschr*<br><br>lüge* | elit*<br>consens*<br>antidemocratic*<br><br>referend*<br>corrot*<br>propagand*<br>politici*<br>ingann*<br><br>tradi*<br><br>vergogn*<br><br>scandal*<br>verità*<br>disonest*<br><br>partitocrazia<br><br><br>menzogn*<br>mentir* |
| <b>Context</b> | establishm*<br>heersend*<br>capitul*<br>kapitul*<br>kaste*<br>leugen*<br>lieg*  | establishm*<br>ruling*   |  |  |

(from Rooduijn and Pauwels 2011)

# Potential disadvantage: Context specific

Lexicons' Accuracy in Document Classification  
Compared to Machine-Learning Approach



**Source:** González-Bailón and Paltoglou (2015)

# How to build a dictionary

- ▶ The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme
- ▶ Three key issues:
  - Validity      Is the dictionary's category scheme valid?
  - Recall        Does this dictionary identify *all* my content?
  - Precision    Does it identify *only* my content?
- ▶ Imagine two logical extremes of including all words (too sensitive), or just one word (too specific)



# How to build a dictionary

1. Identify “extreme texts” with “known” positions. Examples:
  - ▶ Tweets by populist vs mainstream parties (for populism dictionary)
  - ▶ Facebook comments to news about natural catastrophes vs football victories (for sentiment dictionary)
  - ▶ Subreddits for white nationalist groups vs regular politics (for racist rhetoric)
2. Search for differentially occurring words using word frequencies
3. Examine these words in context to check their precision and recall
4. Use regular expressions to see whether stemming or wildcarding is required

## Login details: RStudio Server

RStudio Server URL:

`rstudio.pablobarbera.com`

user = **viennaXX** and password = **passwordXX**

where XX is your assigned number

# Quantitative Text Analysis. Applications to Social Media Research

**Pablo Barberá**

London School of Economics

`www.pablobarbera.com`

Course website:

[pablobarbera.com/text-analysis-vienna](http://pablobarbera.com/text-analysis-vienna)