Computational Political Science

Session 3

David Broska
Zeppelin University
February 16, 2021

Course schedule

Session	Date	Topic	Assignment	Due date
1	Feb 02	Overview and key concepts	-	-
2	Feb 09	Preprocessing and descriptive statistics	Formative	Feb 22 23:59:59
3	Feb 16	**Dictionary methods**	-	-
4	Feb 23	Machine learning (for texts)	Summative 1	Mar 08 23:59:59
5	Mar 02	Supervised scaling models for texts	-	-
6	Mar 09	Unsupervised scaling models for texts	Summative 2	Mar 15 23:59:59
7	Mar 16	Similarity and clustering	-	-
8	Mar 23	Topic models	Summative 3	Apr 12 23:59:59
-	-	*Break*	-	-
9	Apr 13	Retrieving data from the web	-	-
10	Apr 20	Published applications	-	-
11	Apr 27	Project Presentations	-	-

Outline for today

1. **Introduction to dictionary methods**
    - Rationale
    - Dictionaries as classifiers
    - Caveats
2. **Well-known dictionaries**
    - General Inquirer
    - Moral Foundations Dictionary
    - Regressive Imagery Dictionary
    - Linguistic Inquiry & Word Count
3. **Applications**
    - Emotional contagion
    - Policy positions
    - Terrorist speech
4. **How to build a dictionary?**
    - Quality criteria
    - Steps
5. **Coding exercise**

Introduction to dictionary methods

Rationale for dictionaries

Rather than count words that occur, pre-define words associated with specific meanings and count only those

Two components:

1. **key** is the label for the equivalence class for the concept or canonical term
2. **values** are (multiple) terms or patterns that are declared *equivalent occurrences* of the key class
   - Frequently involves stemming/lemmatization of inflected words to capture all relevant terms or patterns

Dictionary vs thesaurus

```r
library(quanteda)
corpus <- c("We aren't schizophrenic but I am",
            "I bought myself a car")
# first person pronouns
fp <- dictionary(list(singular=c("I","me","my","mine","myself"),
                      plural  =c("we","us","our","ourselves")))
dfm(corpus, dictionary = fp)
```

```
## Document-feature matrix of: 2 documents, 2 features (25.0% sparse).
##        features
## docs     singular plural
##   text1         1      1
##   text2         2      0
```

```r
dfm(corpus, thesaurus  = fp)
```

```
## Document-feature matrix of: 2 documents, 9 features (44.4% sparse).
##        features
## docs     SINGULAR PLURAL aren't schizophrenic but am bought a car
##   text1         1      1      1             1   1  1      0 0   0
##   text2         2      0      0             0   0  0      1 1   1
```

Feature weighting

```
( dfmat <- dfm(corpus) )                        # create dfm with counts
```

```
## Document-feature matrix of: 2 documents, 10 features (45.0% sparse).
##        features
## docs    we aren't schizophrenic but i am bought myself a car
##   text1  1      1             1   1 1  1      0      0 0   0
##   text2  0      0             0   0 1  0      1      1 1   1
```

```
( dfmat_w <- dfm_weight(dfmat, scheme = "prop") )   # compute proportion
```

```
## Document-feature matrix of: 2 documents, 10 features (45.0% sparse).
##        features
## docs      we aren't schizophrenic  but    i   am bought myself   a car
##   text1 0.17   0.17          0.17 0.17 0.17 0.17      0      0   0   0
##   text2 0      0             0    0    0.20 0      0.2    0.2 0.2 0.2
```

```
( dfmat_wd <- dfm_lookup(dfmat_w, dictionary = fp) ) # add up relevant cells
```

```
## Document-feature matrix of: 2 documents, 2 features (25.0% sparse).
##        features
## docs    singular plural
##   text1     0.17   0.17
```

Dictionaries as classifiers

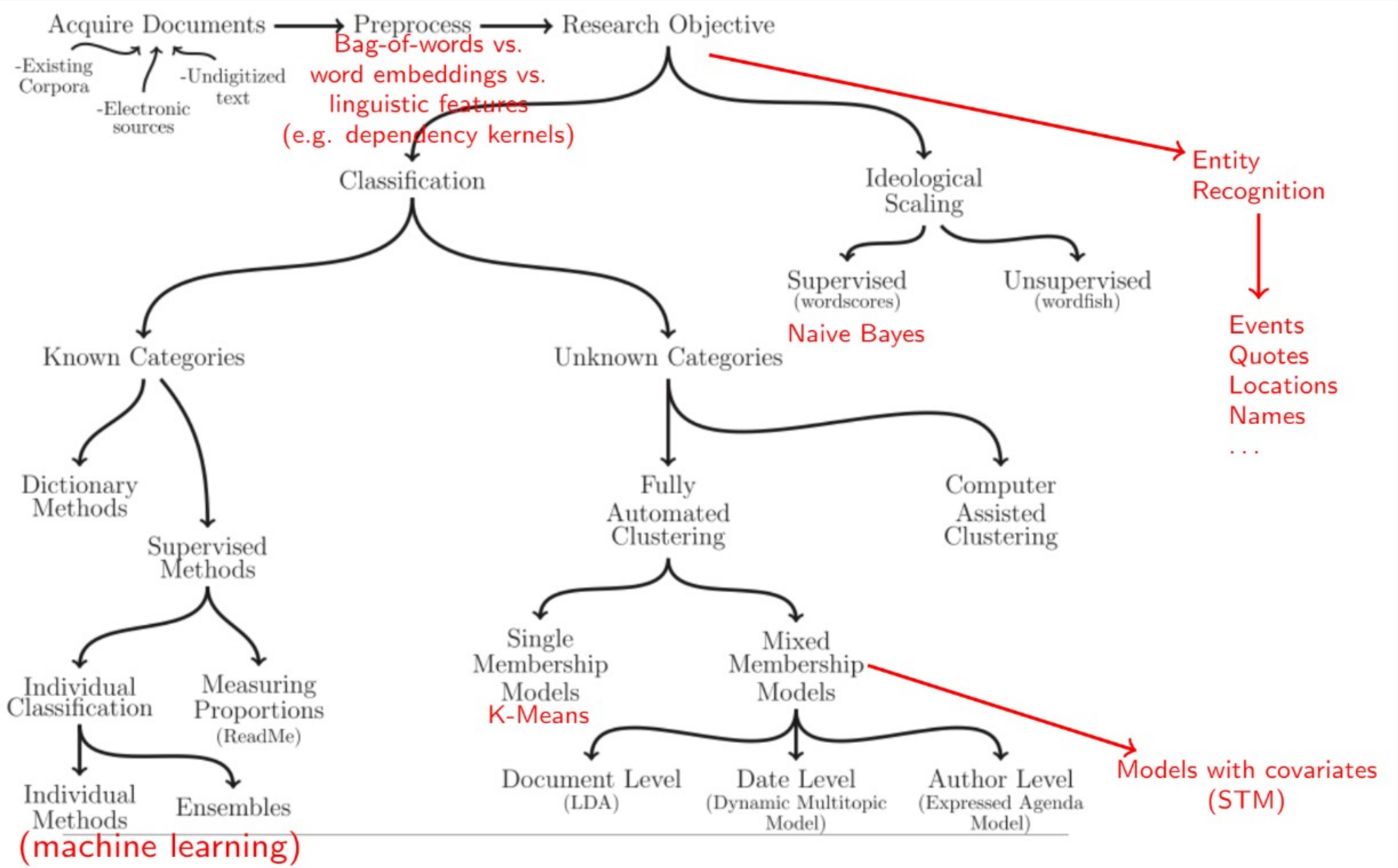


Fig. 1 in Grimmer and Stuart (2013)

Dictionaries as classifiers

Classifying documents into known categories

Lists of words that correspond to each category:

- Emotions: sad, happy, angry, anxious...
- Cognitive processes: Insight, causation, discrepancy, tentative...
- Hate speech: Sexism, homophobia, xenophobia, racism...
- Sentiment: Positive or negative

Count number of times they appear in each document

- Normalize by document length if necessary
- 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

Mixed vs single membership

Mixed membership

A document can belong to more than category

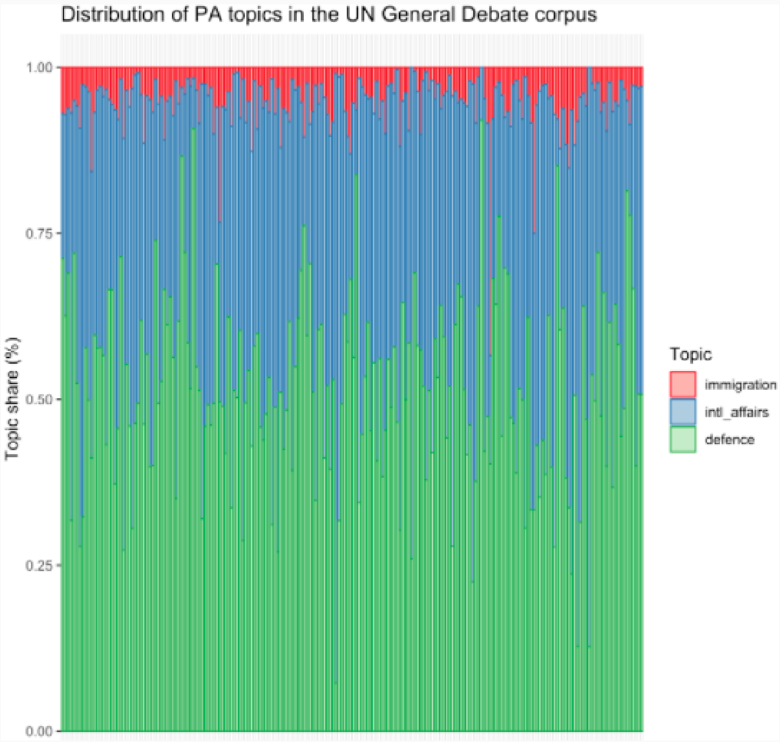
E.g. a speech held at the UN General Debate is about immigration, defense, and other topics.

Single membership

A document belongs to one category

Using dictionaries as single membership classifier requires **simplification**!

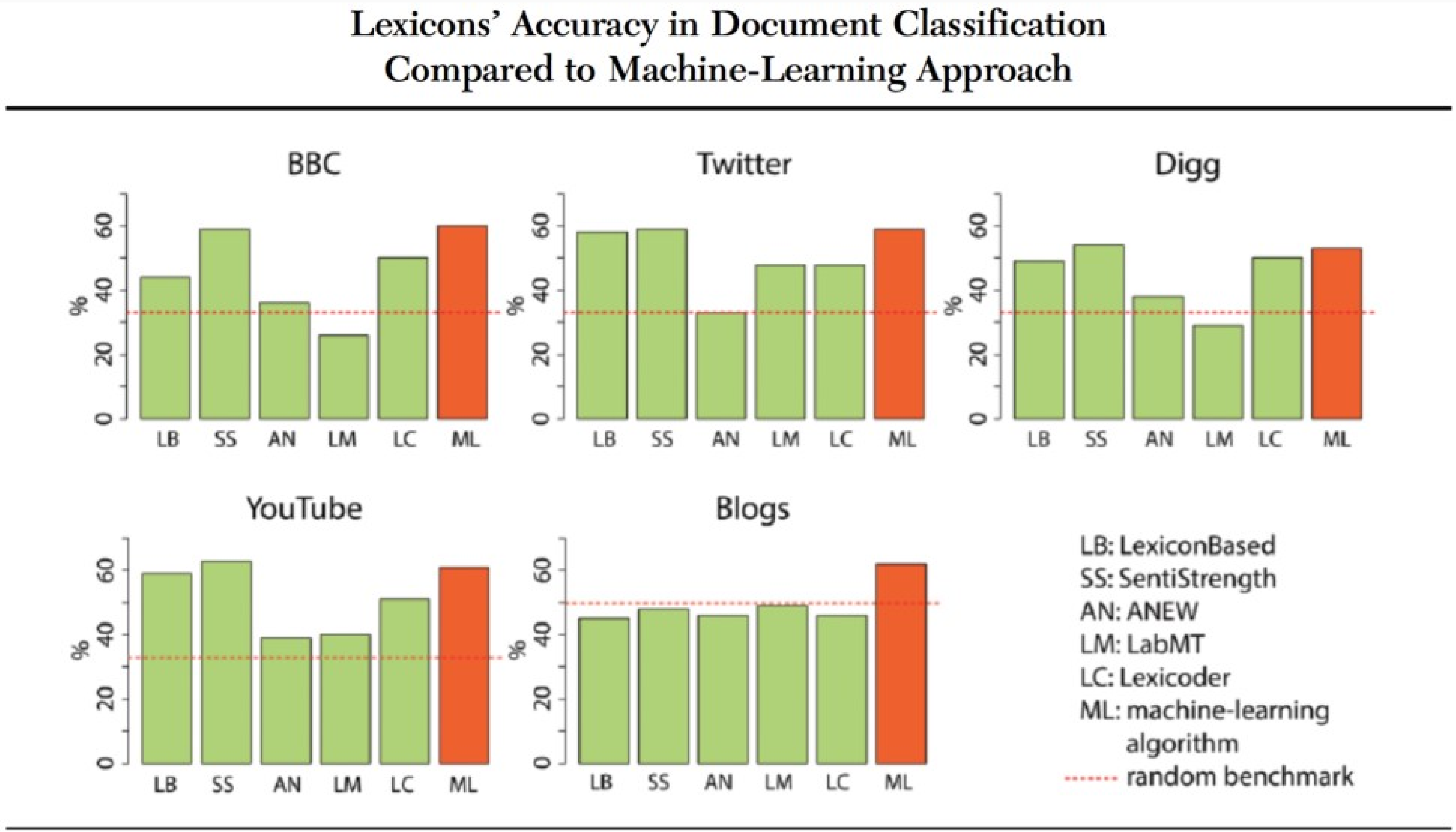
E.g. a document is about defense if the majority of the associated words occur more often than those of the other topics.



Puschmann (2019)

Disadvantage: context specific

Classification accuracy of dictionary methods **depends on the context**



González-Bailón and Paltoglou (2015)

Disadvantage: context specific

Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) file to classify sentiment for a corpus of 50,115 firm-year 10-K filings from 1994-2008

They found two problems with the dictionary approach:

1. Polysemes - words that have multiple meanings

Almost three-fourths of the "negative" words of H4N were typically not negative in a financial context

For example: cost, tax, capital, liability, and vice

2. Incompleteness - dictionary lacked important negative financial words

For example: litigation, restated, misstatement, and unanticipated

Well-known dictionaries

General Inquirer

- Originally developed by Stone et al (1966)
- Latest version contains 182 categories - the "Harvard IV-4" dictionary, the “Lasswell” dictionary, and five categories based on the social cognition work of Semin and Fiedler

Examples

- "self references", containing mostly pronouns
  - self = I, me, my, mine, myself
  - selves = we, us, our, ours, ourselves
- "negatives", the largest category with 2291 entries
  - abandon, fanatical, distract

Also uses simple word sense disambiguation, for example to distinguishes between race as a contest, race as moving rapidly, race as a group of people of common descent, and race in the idiom “rat race”

Output example: http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html

Moral foundations dictionary (MFD)

Definition: Moral foundations are dimensions of human moral reasoning

Moral foundations dictionary by Graham and Haidt:

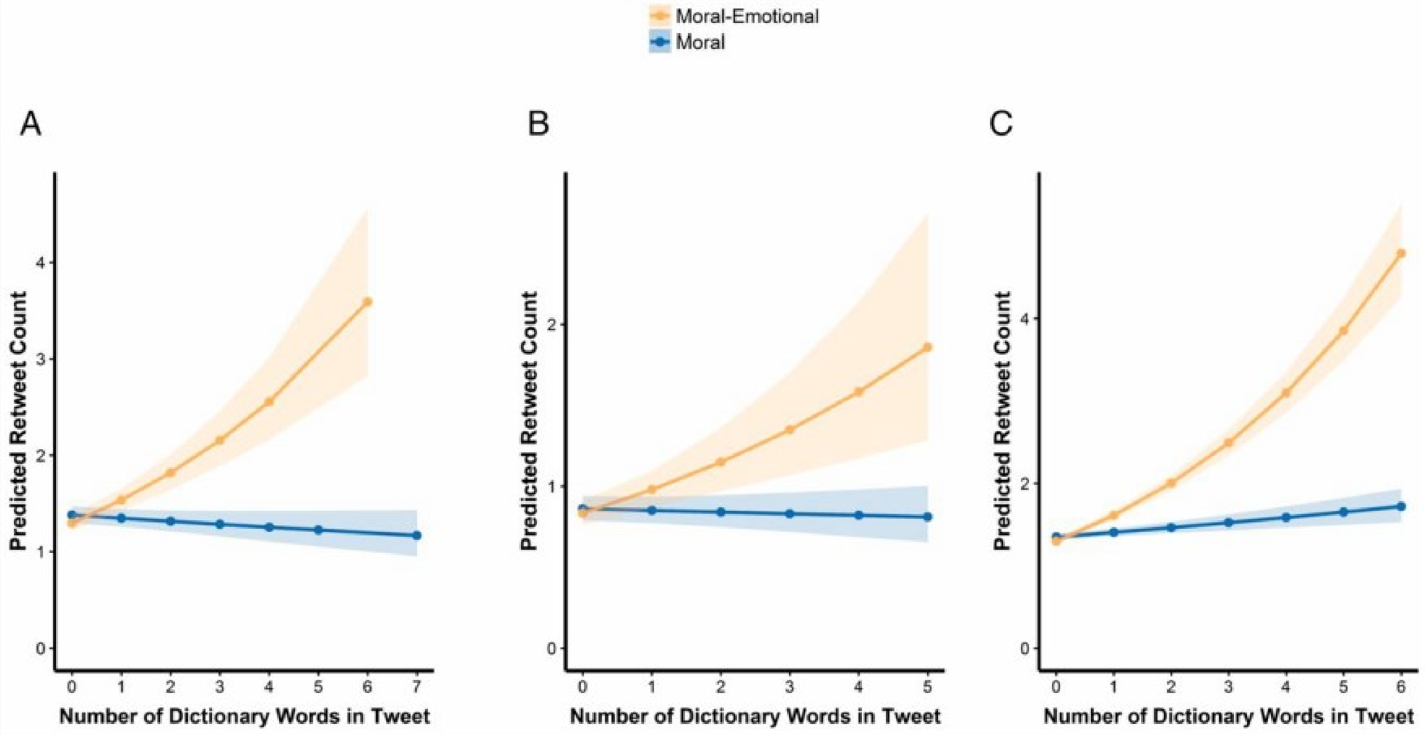
Measures the proportions of virtue and vice words for each *foundation*:

1. *Care/Harm* relates to the ability to feel (and dislike) the pain of others and underlies virtues of kindness, gentleness, and nurturance
2. *Fairness/Cheating* relates to ideas of justice, rights, and autonomy
3. *Loyalty/Betrayal* underlies virtues of patriotism and self-sacrifice for the group
4. *Authority/Subversion* underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions
5. *Sanctity/Degradation* underlies religious notions of striving to live in an elevated, less carnal, more noble way: the body is a temple which can be desecrated by immoral activities

Link: https://www.moralfoundations.org

MFD example

What is the role of moral emotions in the spread of morally tinged posts on Twitter?



Findings of Brady et al (2017)

- Posts with the highest amount of moral-emotional language are retweeted most.
- Moral emotional language increases the diffusion within liberal and conservative clusters and less so across those ideological boundaries

MFD example

Brady et al (2017) use two dictionaries to identify moral and emotional words respectively:

if a certain word occurs in both dictionaries it is treated as a moral-emotional word

Table 1. Sample tweets from each political topic, separated by ideology

Topic	Mean ideology of retweeters	Twitter message
Gun control	Conservative	America needs to Arm itself. Stand and **Fight** for Your Second Amendment Rights. We are literally in a **War** Zone. Carry and get Trained.
	Liberal	Thanks to **greed**, the republication leadership & the #NRA – No one is **safe** #SanBernadino #gunsense #guns #morningjoe
Same-sex marriage	Conservative	Gay marriage is a diabolical, **evil** lie aimed at **destroying** our nation #o4a #news #marriage
	Liberal	New Mormon Policy Bans Children Of Same-Sex Parents-this church wants to **punish** children? Are you kidding me?!? **Shame**
Climate change	Conservative	Leftists take 'global warming' based on **bad** science as **faith** and act on it, but proven voter fraud is just racism #tcot #teaparty
	Liberal	**Fighting** #climatechange is **fighting** hunger. Put your #eyesonParis for a fair climate deal.

Examples of tweets containing at least one moral-emotional word that were retweeted largely by liberals or conservatives. Moral-emotional words are in bold.

Regressive Imagery Dictionary (RID)

- RID is designed to measure primordial vs. conceptual thinking
  - *Conceptual* thought is abstract, logical, reality oriented, and aimed at problem solving
  - *Primordial* thought is associative, concrete, and takes little account of reality - the type of thinking found in fantasy, reverie, and dreams
- Consists of about 3,200 words and roots, assigned to 29 categories of primary process cognition, 7 categories of secondary process cognition, and 7 categories of emotions
- Categories were derived from the theoretical and empirical literature on regressive thought by Martindale (1975, 1990)

Regressive Imagery Dictionary (RID)

Full listing of categories

1 orality	21 brink-passage	41 aggression	62 novelty
2 anality	22 narcissism	42 expressive behaviour	63 negation
3 sex	23 concreteness	43 glory	64 triviality
4 touch	24 ascend	44 female role	65 transmute
5 taste	25 height	45 male fole	
6 odour	26 descent	46 self	
7 general sensation	27 depth	47 related others	
8 sound	28 fire	48 diabolic	
9 vision	29 water	49 aspiration	
10 cold	30 abstract thought	50 angelic	
11 hard	31 social behaviour	51 flowers	
12 soft	32 instrumental behaviour	52 synthesize	
13 passivity	33 restraint	53 streight	
14 voyage	34 order	54 weakness	
15 random movement	35 temporal references	55 good	
16 diffusion	36 moral imperative	56 bad	
17 chaos	37 positive affect	57 activity	
18 unknown	38 anxiety	58 being	
19 timelessness	39 sadness	59 analogy	
20 counscious	40 affection	61 integrative con	

More on categories: http://www.kovcomp.co.uk/wordstat/RID.html

Linguistic Inquiry & Word Count (LIWC)

LIWC reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and parts of speech.

- Hierarchical dictionary which 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 *cried* causes each of these five subdictionary scale scores to be incremented
- Created by James Pennebaker et al - see http://www.liwc.net
- Subject to a small fee: https://liwcsoftware.onfastspring.com
- LIWC is pronounced as Luke

Examples

Emotional contagion

Using the LIWC dictionary, Kramer et al (2014) show that emotional states are transferred to others by exposure to content of Facebook friends

- N=689,003 Facebook users
- Treatment 1: Postive content more visible on news feed
- Treatment 2: Negative content more visible on news feed
- Control: No news feed intervention

Controversial study: concerns about ethics!

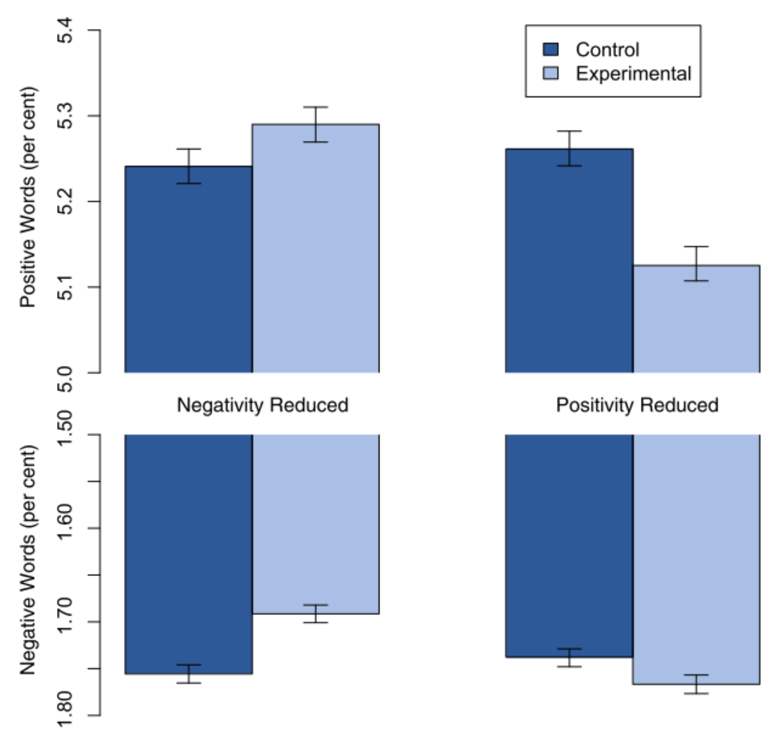


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

Policy positions

- Laver and Garry (2000) created a **hierarchical** set of categories to distinguish policy domains and policy positions
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a general domain
- Looked for word occurrences within "word strings with an average length of ten words"
- Built the dictionary on a set of specific UK manifestos

Table 1 Abridged Section of Revised Manifesto Coding Scheme

1	ECONOMY Role of state in economy
1 1	ECONOMY/+State+ Increase role of state
1 1 1	ECONOMY/+State+/Budget Budget
1 1 1 1	ECONOMY/+State+/Budget/Spending Increase public spending
1 1 1 1 1	ECONOMY/+State+/Budget/Spending/Health
1 1 1 1 2	ECONOMY/+State+/Budget/Spending/Educ. and training
1 1 1 1 3	ECONOMY/+State+/Budget/Spending/Housing
1 1 1 1 4	ECONOMY/+State+/Budget/Spending/Transport
1 1 1 1 5	ECONOMY/+State+/Budget/Spending/Infrastructure
1 1 1 1 6	ECONOMY/+State+/Budget/Spending/Welfare
1 1 1 1 7	ECONOMY/+State+/Budget/Spending/Police
1 1 1 1 8	ECONOMY/+State+/Budget/Spending/Defense
1 1 1 1 9	ECONOMY/+State+/Budget/Spending/Culture
1 1 1 2	ECONOMY/+State+/Budget/Taxes Increase taxes
1 1 1 2 1	ECONOMY/+State+/Budget/Taxes/Income
1 1 1 2 2	ECONOMY/+State+/Budget/Taxes/Payroll
1 1 1 2 3	ECONOMY/+State+/Budget/Taxes/Company
1 1 1 2 4	ECONOMY/+State+/Budget/Taxes/Sales
1 1 1 2 5	ECONOMY/+State+/Budget/Taxes/Capital
1 1 1 2 6	ECONOMY/+State+/Budget/Taxes/Capital gains
1 1 1 3	ECONOMY/+State+/Budget/Deficit Increase budget deficit
1 1 1 3 1	ECONOMY/+State+/Budget/Deficit/Borrow
1 1 1 3 2	ECONOMY/+State+/Budget/Deficit/Inflation

Terrorist speech

Analysis of Al Qaeda discourse in videotapes, interviews, and letters by Hancock et al (2010)

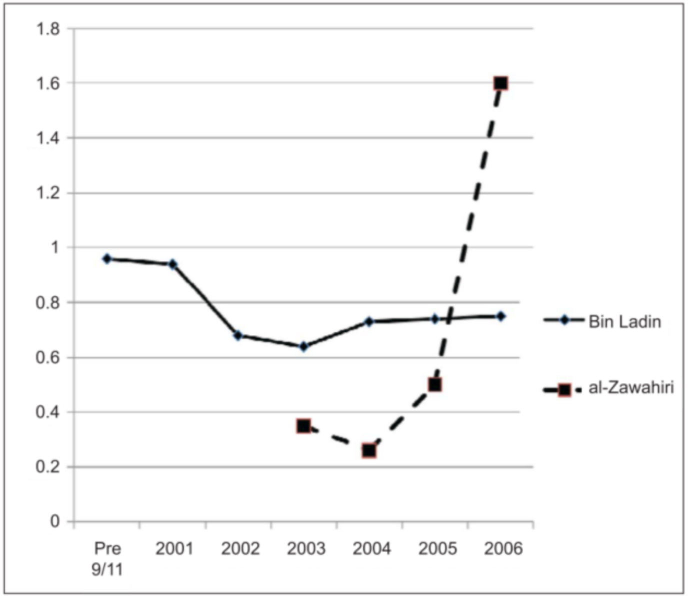


Figure 1. First person singular pronoun use by bin Laden and Al Zawahiri pre-9/11 to 2006.

First-person pronouns (I, me, my, mine) included in LIWC:

- Osama bin Laden's use remained constant over time
- Ayman al-Zawahri increased usage over time

Suggests: Zawahiri was feeling threatened, indicating a rift in his relationship with bin Laden

Terrorist speech

Using the LIWC dictionary to extract grammatical pronouns and words related to cognitive mechanisms, Hancock et al (2010) find that **high status** organization members use

- fewer words writing to lower status
- fewer first person singular pronouns than lower status members
- "you" significantly more often
- far fewer cognitive mechanism words (indicating cause, discrepancy and inclusion)

	High-low	**Same status**	**Low-high**
Word count	63.85 (10.07)	88.70 (19.62)	187.95 (25.70)
'I'	0 (0)	0.55 (0.29)	0.38 (0.23)
'You'	2.20 (0.48)	1.32 (0.28)	0.38 (0.16)
Cognitive mechanisms	11.56 (0.73)	14.36 (0.96)	14.18 (0.97)

Table: Mean and standard errors of language features for high-status to low-status messages, same status messages, low-status to high status messages

Terrorist speech

Table 1 Comparison of Public Statements by bin Laden, al-Zawahiri, and Other Terrorist Groups

	bin Laden (1988–2006) (n = 28)†	*al-Zawahiri (2003–2006) (n = 15)*†	*Controls (n = 17)*	*(2-Tailed) p ≤*
Word count	2511.5††	1996.4	4767.5	
Big words (greater than 6 letters)	21.2_{a}†††	23.6_{b}	21.1_{a}	.05
Pronouns	9.15_{ab}	9.83_{b}	8.16_{a}	.09
I (e.g., I, me, my)	0.61	0.90	0.83	
We (e.g., we, our, us)	1.94	1.79	1.95	
You (e.g., you, your, yours)	1.73	1.69	0.87	
He/she (e.g., he, hers)	1.42	1.42	1.37	
They (e.g., they, them)	2.17_{a}	2.29_{a}	1.43_{b}	.03
Propositions	14.8	14.7	15.0	
Articles (e.g., a, an, the)	9.07	8.53	9.19	
Exclusive words (e.g., but, exclude)	2.72	2.62	3.17	
Affect	5.13_{a}	5.12_{a}	3.91_{b}	.01
Positive emotion (e.g., happy, joy, love)	2.57_{a}	2.83_{a}	2.03_{b}	.01
Negative emotion (e.g., awful, cry, hate)	2.52_{a}	2.28_{ab}	1.87_{b}	.03
Anger words (e.g., hate, kill)	1.49_{a}	1.32_{a}	0.89_{b}	.01
Cognitive mechanisms	4.43	4.56	4.86	
Time (e.g., clock, hour)	2.40_{b}	1.89_{a}	2.69_{b}	.01
Past tense verbs	2.21_{a}	1.63_{a}	2.94_{b}	.01
Social processes	11.4_{a}	10.7_{ab}	9.29_{b}	.04
Humans (e.g., child, people, selves)	0.95_{ab}	0.52_{a}	1.12_{b}	.05
Family (e.g., mother, father)	0.46_{ab}	0.52_{a}	0.25_{b}	.08
Content				
Death (e.g., dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g., buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g., faith, Jew, sacred)	2.41	1.84	1.89	

Table 1 in Pennebaker and Chung (2007) based on the LIWC dictionary

"Striking difference between other extremist groups and the two Al-Qaeda authors"

- More focus more on other individuals: “the group is defining itself to a large degree by the existence of an oppositional group" (third-person plural pronouns)
- More emotional statements: "far more emotional in their use of both positive and negative emotion words"
- More anger and hostility words (relative to anxiety or sadness words).

How to build a dictionary

Dictionary: quality criteria

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?

Say we want to classify texts into positive and negative classes:

- What if we included the word *terribly*? For instance, *terribly happy*
- What if we included only the word *distraught*?
- What if we included every word used in the corpus?

Dictionary construction

Steps

1. Identify "extreme texts" with "known" positions. Examples:
   - Tweets by populist vs mainstream parties (for populism dictionary)
   - Opposition leader and Prime Minister in a no-confidence debate (for opposition vs government 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 using wildcards is required

Coding exercise

References

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. "Emotion Shapes the Diffusion of Moralized Content in Social Networks." Proceedings of the National Academy of Sciences 114 (28): 7313–18. https://doi.org/10.1073/pnas.1618923114.

GONZÁLEZ-BAILÓN, SANDRA, and GEORGIOS PALTOGLOU. 2015. "Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources." The Annals of the American Academy of Political and Social Science 659: 95–107.

Hancock, Jeffrey T., David I. Beaver, Cindy K. Chung, Joey Frazee, James W. Pennebaker, Art Graesser, and Zhiqiang Cai. 2010. "Social Language Processing: A Framework for Analyzing the Communication of Terrorists and Authoritarian Regimes." Behavioral Sciences of Terrorism and Political Aggression 2 (2): 108–32. https://doi.org/10.1080/19434471003597415.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." Proceedings of the National Academy of Sciences 111 (24): 8788–90. https://doi.org/10.1073/pnas.1320040111.

Pennebaker, James W., and Cindy K. Chung. 2007. "Computerized Text Analysis of Al-Qaeda Transcripts." In A Content Analysis Reader, edited by Klaus Krippendorff and M. Bock. CA: SAGE.