Quantitative Text Analysis – Essex Summer School

Dictionaries

dr. Martijn Schoonvelde

University of Groningen

Today's class

Categorize text:

- Pros and cons of using off-the-shelf dictionaries
- Developing a dictionary
- Lab session

Analysis

- Three broad types of analysis (Boumans & Trilling 2016), from most deductive to most inductive:
    - **counting and dictionary methods**: the researcher can fully specify relevant features, and will categorise text accordingly
    - **supervised methods**: the researcher knows how to categorise documents, and uses machine learning methods to learn which features drive this categorisation
    - **unsupervised methods**: the researcher uses qta tools to learn about textual categorisation inductively

Dictionary methods

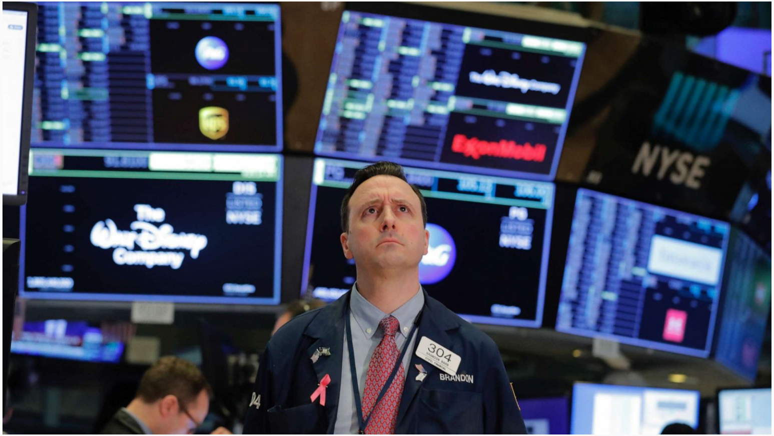
- “Dictionaries use the rate at which key words appear in a text to classify documents into categories or to measure the extent to which documents belong to particular categories” (Grimmer & Stewart, 2013)
  - That is, count words associated with specific meanings
- Dictionaries consist of key-value pairs
  - Key: label for the concept or equivalence class
  - Values: (multiple) terms or patterns of terms that are declared equivalent occurrences of the key class

Dictionary methods

Under what conditions do dictionary methods excel? (Van Atteveldt *et al.* 2022)

- The categories that we want to code are manifest and concrete rather than latent and abstract
  - Abstract theoretical constructs such as frames or ideologies are difficult to capture using dictionaries
- All known synonyms are included in the dictionary
- Dictionary entries do not have multiple meanings

Dictionaries and the economy



Source https://www.inversorglobal.com/2020/11/los-mejores-brokers-argentinos-de-2020/

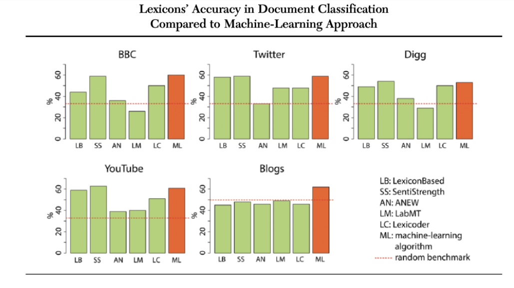
Dictionary methods

Key questions when applying dictionaries concern their validity, recall, and precision

- Validity – does the dictionary meaningfully operationalize our concept of interest?
- Recall – does the dictionary identify all relevant content?
- Precision – does the dictionary identify only relevant content?

Dictionary methods

Off-the-shelf dictionaries have a bit of a bad reputation. For example, lots of research that shows that sentiment dictionaries vary in their performance across types of text



Source: González Bailón & Paltoglu 2015

Issues with Dictionary Methods

Domain-specificity problem (Chan *et al.* 2020)

- No context – word can have multiple meanings (polysemous words)
- Domain specific
  - "honourable" in HoC speeches
- Unknown unknowns
  - False negatives
- English bias in dictionary construction – do they translate to other languages?

Recent advancements

Recent studies have made progress on these issue by developing domain-specific dictionaries that rely on machine translation, word embeddings or extensive human coding (Müller, 2021; Proksch *et al.*, 2019; Rauh, 2018; Rheault *et al., 2016*; van Atteveldt *et al.*, 2008; Widmann, 2021).

Also: joint modeling of topics and sentiment – Joint Sentiment Topic model (JST) and the reversed Joint Sentiment Topic model (rJST) (Lin *et al.*, 2009; Lin *et al.*, 2012).

NRC Word-Emotion Association Lexicon

- Mohammad and Turney (2013)
- Annotation crowdsourced through Mechanical Turk
- NRC available through the package **quanteda.sentiment**
- Available in multiple languages

```
> summary(data_dictionary_NRC)
             Length Class  Mode
anger        1247   -none- character
anticipation  839   -none- character
disgust      1058   -none- character
fear         1476   -none- character
joy           689   -none- character
negative     3324   -none- character
positive     2312   -none- character
sadness      1191   -none- character
surprise      534   -none- character
trust        1231   -none- character
```

Lexicoder Sentiment Dictionary

- Young and Soroka (2012)
- Valence dictionary, which builds on earlier dictionaries: Roget's Thesaurus, General Inquirer and the Regressive Image Dictionary
  - Positive: if $+++$ or if $++$ *NA*;
  - Negative if $---$ or if $--$ *NA*
  - For other words, contextual analysis through KWIC and other preprocessing steps to account for ambiguity
- 4567 positive and negative words
- Validated against 3 coders of 900 *New York Times* articles and other dictionaries

Lexicoder Sentiment Dictionary

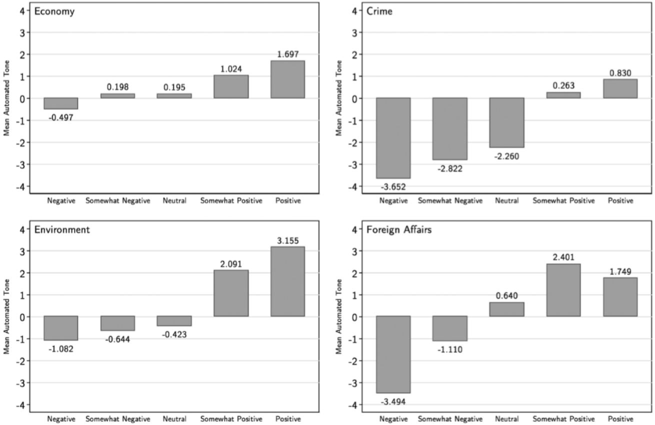


Figure 2. Comparing results across topics.

- Trend in line with expetactions
- But levels differ across topics

LIWC

LIWC			LIWC Cont.		
Category	**Example**	**T-statistics**	**Category**	**Example**	**T-statistics**
Linguistics Processes			Negative emotion	hurt, ugly, nasty	6.49***
Words > 6 letters		-3.41**	Anxiety	fearful, nervous	2.37
Dictionary words		9.60****	Anger	hate, kill, annoy	5.30***
Total function words		8.98****	Sadness	cry, grief, sad	3.54***
Personal pron.	I, them, her	7.07****	Cognitive process	cause, ought	6.09***
1st pers singular	I, me, mine	9.83****	Insight	think, know	0.11
1st pers plural	we, us, our	-2.38	Causation	effect, hence	0.93
2nd person	you, your, thou	-0.91	Discrepancy	should, would	5.53***
3rd pers singular	she, her, him	3.63**	Tentative	maybe, perhaps	5.95***
3rd pers plural	their, they'd	2.47	Certainty	always, never	4.02***
Impersonal pron.	it, it's, those	7.07****	Inhibition	block, constrain	0.32
Articles	a, an, the	4.13***	Inclusive	with, include	4.74 ***
Common verbs	walk, went, see	6.27***	Exclusive	but, without	7.53 ****
Auxiliary verbs	am, will, have	5.76***	Perceptual process		1.93
Past tense	went, ran, had	8.70****	See	view, saw, seen	1.68
Present tense	is, does, hear	4.00***	Hear	listen, hearing	-0.88
Future tense	will, gonna	5.84***	Feel	feels, touch	1.94
Adverbs	very, really	7.92****	Biological process		4.22***
Prepositions	to, with, above	7.62****	Body	cheek, spit	5.02***
Conjunctions	and, whereas	4.59***	Health	clinic, flu, pill	1.51
Negations	no, not, never	1.71	Sexual	horny, incest	-0.61
Quantifiers	few, many, much	2.98*	Ingestion	dish, eat, pizza	4.37***
Numbers	second, thousand	-3.68**	Relativity	area, bend, exit	9.52 ****
Swear words	damn, piss, fuck	5.53***	Motion	arrive, car	3.07*
Spoken Categories			Space	down, in, thin	8.87****
Assent	agree, OK, yes	7.05****	Time	end, until	5.87***
Nonfluency	er, hm, umm	1.41	**Personal Concerns**		
Filters	blah, imean		Work	job, majors	0.05
Psychological			Leisure	chat, movie	2.97*
Social process	mate, talk, child	0.10	Achievement	earn, win	-1.22
Family	son, mom, aunt	2.24	Home	family, kitchen	3.37**
Friends	buddy, neighbor	2.10	Money	audit, cash	0.23
Humans	adult, baby, boy	0.89	Religion	church, altar	-0.77
Affective process	happy, cry	3.55**	Death	bury, coffin	0.49
Positive emotion	love, nice, sweet	0.08			

Table 1. Two-sample T-test statistics of linguistic variables between geo-locator and non-locators. Significant differences of each LIWC attribute are indicated in the third column. (*p <0.01, **p<0.001, ***p<0.0001, ****p<1e-10)

- Developed by James Pennebaker and colleagues
- Extensively validated on various corpora

Dictionary methods in quanteda

- A dictionary in **quanteda** assigns possible features to categories, or "keys". It can have as many categories as you wish, and can be composed of any possible feature.

```
> txt_dfm <- corpus(c("this is excellent", "bad", "good, not horrible")) %>%
tokens() %>% dfm()
> sent_dict <- dictionary(list(positive=c("great","good","excellent"),
+                               negative=c("bad", "horrible", "badly")))
> dfm_dict <- dfm_lookup(txt_dfm, dictionary = sent_dict)
> dfm_dict
Document-feature matrix of: 3 documents, 2 features (33.33% sparse) and 0 docvars
       features
docs    positive negative
  text1        1        0
  text2        0        1
  text3        1        1
```

Sentiment dictionaries in quanteda

Through **quanteda.textmodels** you have access to a number of off-the-shelf sentiment dictionaries:

- Polarity dictionaries have two lists of words, each indicating one "pole" (by default "positive" and "negative")
- 'Valence dictionaries have continuous values/weights associated with each word in a given category, and may have more or fewer than two categories.

Name	Description	Polarity	Valence
data_dictionary_AFINN	Nielsen's (2011) 'new ANEW' valenced word list		✔
data_dictionary_ANEW	Affective Norms for English Words (ANEW)		✔
data_dictionary_geninqposneg	Augmented General Inquirer *Positiv* and *Negativ* dictionary	✔	
data_dictionary_HuLiu	Positive and negative words from Hu and Liu (2004)	✔	
data_dictionary_LoughranMcDonald	Loughran and McDonald Sentiment Word Lists	✔	
data_dictionary_LSD2015	Lexicoder Sentiment Dictionary (2015)	✔	
data_dictionary_NRC	NRC Word-Emotion Association Lexicon	✔	
data_dictionary_Rauh	Rauh's German Political Sentiment Dictionary	✔	
data_dictionary_sentiws	SentimentWortschatz (SentiWS)	✔	✔

Source: `https://github.com/quanteda/quanteda.sentiment/`

Do's when using off-the-shelf dictionaries

- Read the dictionary
  - Domain-specific yes or no?
  - Change the dictionary if necessary – but be transparent about it (i.e., report it)
- Try out on a subset of your corpus – does it pick up on the construct you try to capture? And only your construct?
  - Assess specificity and sensitivity

Developing Dictionaries

- "When counting is automated, success or failure largely reflects a correct (excluding irrelevant data) and exhaustive (including all the relevant data)" (Boumans and Trilling 2016)
- As always, there is no silver bullet approach – much depends on what construct the dictionary is supposed to capture

Developing Dictionaries

Van Atteveldt *et al.* (2022) propose the following workflow for constructing a dictionary:

1. Construct a dictionary based on theoretical considerations and by closely reading a sample of example texts.
2. Code some articles manually and compare with the automated coding.
3. Improve your dictionary and check again.
4. Manually code a validation dataset of sufficient size. The required size depends a bit on how balanced your data is – if one code occurs very infrequently, you will need more data.
5. Calculate the agreement.
   - Precision and recall

Developing Dictionaries

1. Identify "extreme texts" with "known" categories / positions
    - Opposition leader and Prime Minister in a no-confidence debate
    - Five-star review of a product (excellent) and a one-star review (terrible)
2. Search for differentially occurring words using word frequencies
3. Examine these words in context to check their sensitivity and specificity
4. Examine inflected forms to see whether stemming or wildcarding is required
5. Use these words (or their lemmas) for categories

`Source: https://lse-my459.github.io/`

Assessing validity

Content validation

- Check sensitivity of results to exclusion of specific words.
- Disambiguate. For exampkle, check grammatical function of ambiguous words using POS tagging.

Concept validation

- Code a few documents manually and see if dictionary prediction aligns with human coding of document (check precision and recall)
- Check if categorisation behaves in predictable ways. For example, do newspaper articles about the economy peak during periods of economic uncertainty