Dictionary based text classification

Mate Akos

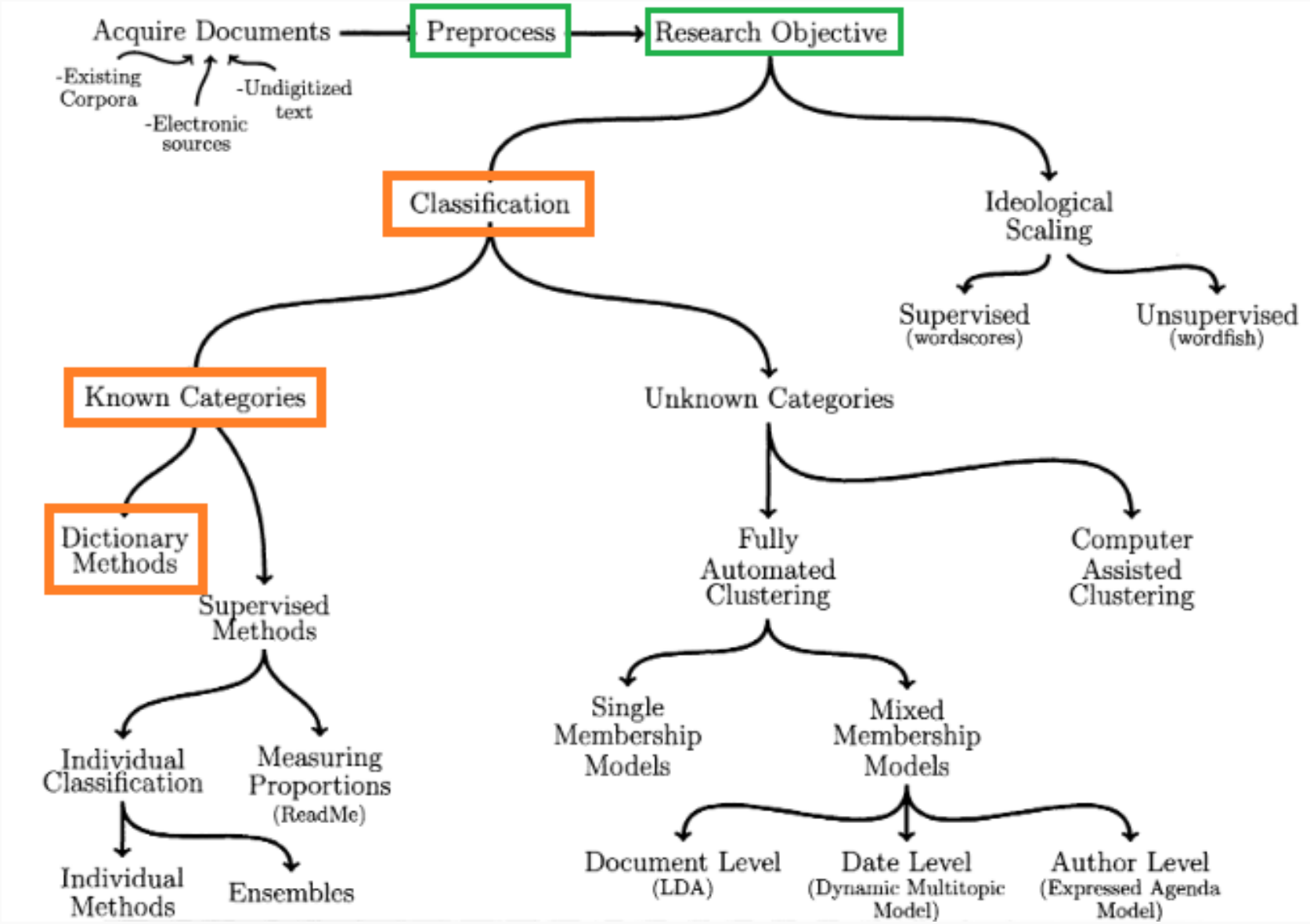
2019 September

MTA TK PTI

For Today

1. Where are we?
2. Dictionary based classification
3. Pros and cons
4. Some notable dictionaries
5. Building a dictionary

Dictionary methods



Dictionaries

- ”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 and Stewart 2013, 274)

Dictionaries

- ”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 and Stewart 2013, 274)
- Nonstatistical (frequency or categorical) analysis. ”It involves counting the frequency of definitive keywords in a text.” (Young and Soroka 2012, 208)

Dictionaries

- ”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 and Stewart 2013, 274)
- Nonstatistical (frequency or categorical) analysis. ”It involves counting the frequency of definitive keywords in a text.” (Young and Soroka 2012, 208)
- It is a set of words that aim to capture a given category.

Dictionaries

- ”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 and Stewart 2013, 274)
- Nonstatistical (frequency or categorical) analysis. ”It involves counting the frequency of definitive keywords in a text.” (Young and Soroka 2012, 208)
- It is a set of words that aim to capture a given category.
- Dictionary building involves a great deal of qualitative assessment

Dictionaries

- ”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 and Stewart 2013, 274)
- Nonstatistical (frequency or categorical) analysis. ”It involves counting the frequency of definitive keywords in a text.” (Young and Soroka 2012, 208)
- It is a set of words that aim to capture a given category.
- Dictionary building involves a great deal of qualitative assessment
- Validate, validate, validate

Dictionaries, how do they look like?

Example dictionary for negative and positive categories (random items from the Lexicoder dictionary):

- **negative:** ”nonsens*”, ”vengeanc*”, ”ironic*”, ”unpredictab*”, ”shamefaced*”
- **positive:** ”high regard”, ”hardi*”, ”goodx”, ”trusts”, ”spectacular*”

Pros and cons

Advantages of the dictionary method

- Perfect reliability (vs intercoder reliability problems)

Pros and cons

Advantages of the dictionary method

- Perfect reliability (vs intercoder reliability problems)
- They are able to capture the latent dimensions in texts

Pros and cons

Advantages of the dictionary method

- Perfect reliability (vs intercoder reliability problems)
- They are able to capture the latent dimensions in texts
- No priors affecting coding judgement

Pros and cons

Advantages of the dictionary method

- Perfect reliability (vs intercoder reliability problems)
- They are able to capture the latent dimensions in texts
- No priors affecting coding judgement
- wide range of applicability (measuring populism, racism, emotions)

Pros and cons

Advantages of the dictionary method

- Perfect reliability (vs intercoder reliability problems)
- They are able to capture the latent dimensions in texts
- No priors affecting coding judgement
- wide range of applicability (measuring populism, racism, emotions)
- Can be adapted for other languages relatively cheaply (e.g.: using Google Translate)

Possible drawbacks of dictionary based frequency analysis

- Effectiveness depends on the match between dictionary and domain of the documents

Possible drawbacks of dictionary based frequency analysis

- Effectiveness depends on the match between dictionary and domain of the documents
- bag-of-words approach: context of words might be lost (e.g.: race, tax, etc.)

Possible drawbacks of dictionary based frequency analysis

- Effectiveness depends on the match between dictionary and domain of the documents
- bag-of-words approach: context of words might be lost (e.g.: race, tax, etc.)
- Dictionary building is a qualitative process (see KWIC in Yound and Soroka (2012) and tax example in Laver and Garry (2000))

Possible drawbacks of dictionary based frequency analysis

- Effectiveness depends on the match between dictionary and domain of the documents
- bag-of-words approach: context of words might be lost (e.g.: race, tax, etc.)
- Dictionary building is a qualitative process (see KWIC in Yound and Soroka (2012) and tax example in Laver and Garry (2000))
- Creating a good dictionary is costly

The problem of context

The key problem with dictionaries not measuring what they intend to: **polysemes**. See Loughran and McDonald (2011) in the extra readings.

- Applying the Harvard-IV-4 dictionary to financial corpus gave wrong results

The problem of context

The key problem with dictionaries not measuring what they intend to: **polysemes**. See Loughran and McDonald (2011) in the extra readings.

- Applying the Harvard-IV-4 dictionary to financial corpus gave wrong results
- general sentiments do not apply to specialized language (e.g.: *vice*, *tax*, *profit*, *liability* are not negative words in a financial context)

The problem of context

The key problem with dictionaries not measuring what they intend to: **polysemes**. See Loughran and McDonald (2011) in the extra readings.

- Applying the Harvard-IV-4 dictionary to financial corpus gave wrong results
- general sentiments do not apply to specialized language (e.g.: *vice*, *tax*, *profit*, *liability* are not negative words in a financial context)
- Some items are not even present in the dictionary (missing the latent dimension entirely)

The problem of context

The key problem with dictionaries not measuring what they intend to: **polysemes**. See Loughran and McDonald (2011) in the extra readings.

- Applying the Harvard-IV-4 dictionary to financial corpus gave wrong results
- general sentiments do not apply to specialized language (e.g.: *vice*, *tax*, *profit*, *liability* are not negative words in a financial context)
- Some items are not even present in the dictionary (missing the latent dimension entirely)
- Extreme word frequency issues might bias results (Pury 2011 in the extra readings)

Table 2

Pairwise correlations, automated dictionaries

	LSD	GI	ROG	RID	ANEW	DAL	LIWC	PMI	TAS/C
GI	0.672								
ROG	0.471	0.469							
RID	0.669	0.480	0.350						
ANEW	0.500	0.464	0.236	0.367					
DAL	0.519	0.481	0.285	0.385	0.482				
LIWC	0.753	0.598	0.428	0.663	0.488	0.490			
PMI	0.228	0.172	0.093	0.128	0.115	0.201	0.159		
TAS/C	0.663	0.601	0.455	0.513	0.438	0.432	0.635	0.178	
WNA	0.230	0.220	0.102	0.068	0.076	0.155	0.224	0.176	0.178

Note. $N = 900$. All correlations are significant at $p < .001$.

Dictionaries and preprocessing

- Check the dictionary, does it include:
  - lowercase / uppercase
  - wordstems / wildcards
  - unigrams or bi or tri-grams?
- our dfm needs to match the dictionary format to get matches

Some dictionaries

- General Inquirer (GI)
- Lexicoder Sentiment Dictionary (LSD)
- Linguistic Inquiry and Word Count (LIWC)
- NRC Word-Emotion Lexicon

General Inquirer

- Introduced in Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.
- Currently: containts the ”Harvard IV-4” and the ”Lasswell” dictionaries
- 182 categories (largest is ”negative” with 2291 entry)
- more info: `http://www.wjh.harvard.edu/~inquirer/homecat.htm`

Lexicoder sentiment dictionary

- Two categories: positive; negative
- Domain: political texts (media)
- Composed from Roget's Thesaurus; GI; and the Regressive Imagery Dictionary (RID)
- Method: if word is POS in three source $\rightarrow$ POS (or POS in two source with NA in third)

Categories in LSD (some attention to context)

- negative (2858 items)
- positive (1709 items)
- positive words preceded by negation (1721)
- negative words preceded by negation (2860)

LIWC

- It is a commercial software
- ”it calculates the percentage of total words that match each of the dictionary categories”
- 82 categories and 4500 words and wordstems
- Developed by Pennebaker et al.
- More info: Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1), 24-54.

(Open source alternative to LIWC: VADER `https://github.com/cjhutto/vaderSentiment`)

NRC Word-Emotion Lexicon

Summary Details of the NRC Emotion Lexicon

Association Lexicon	Version	# of Terms	Categories	Association Scores	Method of Creation	Papers
Word-Emotion and Word-Sentiment Association Lexicon						
NRC Word-Emotion Association Lexicon (also called EmoLex) README	0.92 (2010)	14,182 unigrams (words) ~25,000 senses*	**sentiments:** negative, positive **emotions:** anger, anticipation, disgust, fear, joy, sadness, surprise, trust	0 (not associated) or 1 (associated) not associated, weakly, moderately, or strongly associated	**Manual:** By crowdsourcing on Mechanical Turk. **Domain:** General	**Crowdsourcing a Word-Emotion Association Lexicon**, Saif Mohammad and Peter Turney, *Computational Intelligence*, 29 (3), 436-465, 2013. Paper (pdf) BibTeX **Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon**, Saif Mohammad and Peter Turney, In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, June 2010, LA, California. Abstract Paper (pdf) Presentation

Dictionary building

- ”[...] a good quantitative content analysis dictionary will consist of words with as little ambiguity as possible.” (Laver and Garry 2000, 625)

Dictionary building

- ”[...] a good quantitative content analysis dictionary will consist of words with as little ambiguity as possible.” (Laver and Garry 2000, 625)
- Goal: have all the relevant words in the given category (perfect validity)

Dictionary building

- ”[...] a good quantitative content analysis dictionary will consist of words with as little ambiguity as possible.” (Laver and Garry 2000, 625)
- Goal: have all the relevant words in the given category (perfect validity)
- Think hard about

Dictionary building

- ”[...] a good quantitative content analysis dictionary will consist of words with as little ambiguity as possible.” (Laver and Garry 2000, 625)
- Goal: have all the relevant words in the given category (perfect validity)
- Think hard about
  - validity: are we capturing the categories we want with the dictionary?
  - precision: $\frac{TP}{TP+FP}$ How well the dictionary identifies the category only?
  - recall (or sensitivity): $\frac{TP}{TN+FP}$ How much of the content is miscategorized?

Confusion matrix for precision-recall

		Predicted	
		cat1	cat0
Actual	cat1	TP	FN
	cat0	FP	TN

- TP for words that are class 1 and predicted in class 1
- FN for words that are class 1 and predicted in class 0
- FP for words that are class 0 and predicted in class 1
- TN for words that are class 0 and predicted in class 0

Strategies for building dictionaries

- Deductive (theory driven): content validity is important!
- Inductive: using reference texts

Strategies for building dictionaries

- Deductive (theory driven): content validity is important!
- Inductive: using reference texts
  - use texts with known positions to identify words (party manifestos of extreme right/left)
  - use KWIC to see context (precision, recall)
  - use word frequecies to see influental terms
  - make a decision if stemming or wildcarding would be used

Validation

Validation is key for dictionary based analysis

- Face validity: the output of the dictionary looks the way we would expect (based on some theory or argument)

Validation

Validation is key for dictionary based analysis

- Face validity: the output of the dictionary looks the way we would expect (based on some theory or argument)
- Concurrent validity: our dictionary correlates with another pre-existing one, measuring the same concept

Validation

Validation is key for dictionary based analysis

- Face validity: the output of the dictionary looks the way we would expect (based on some theory or argument)
- Concurrent validity: our dictionary correlates with another pre-existing one, measuring the same concept
- Predictive validity: evaluate the predictive capabilities of our dictionary against a predicted outcome

Validation

Validation is key for dictionary based analysis

- Face validity: the output of the dictionary looks the way we would expect (based on some theory or argument)
- Concurrent validity: our dictionary correlates with another pre-existing one, measuring the same concept
- Predictive validity: evaluate the predictive capabilities of our dictionary against a predicted outcome
- Construct validity: are we actually measuring the concept that we aim to?