Text as Data: Dictionaries

Guest Course – January 2026

Germain Gauthier, Philine Widmer[1]

[1]Bocconi Unversity, Paris School of Economics

USI Lugano

Overview of dictionary-based methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
- They can be corpus-specific: counting sets of words or phrases across documents
  - e.g., number of times a judge says "justice" vs. "efficiency"
- Or more general dictionaries:
  - e.g., WordNet, LIWC, NRC Emotion Lexicon

WordNet

- English word database: 118K nouns, 12K verbs, 22K adjectives, 5K adverbs

The noun "bass" has 8 senses in WordNet.
1. bass^{1} - (the lowest part of the musical range)
2. bass^{2}, bass part^{1} - (the lowest part in polyphonic music)
3. bass^{3}, basso^{1} - (an adult male singer with the lowest voice)
4. sea bass^{1}, bass^{4} - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass^{1}, bass^{5} - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass^{6}, bass voice^{1}, basso^{2} - (the lowest adult male singing voice)
7. bass^{7} - (the member with the lowest range of a family of musical instruments)
8. bass^{8} - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Figure 19.1 A portion of the WordNet 3.0 entry for the noun *bass*.

- Synonym sets (synsets) are a group of near-synonyms, plus a gloss (definition)—also contains information on antonyms (opposites)
- Nouns are organized in a categorical hierarchy (hence "WordNet")
  - "hypernym" – the higher category that a word is a member of
  - "hyponyms" – members of the category identified by a word

WordNet supersenses (word categories)

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure 19.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

Supersense	Verbs denoting ...
body	grooming, dressing and bodily care
change	size, temperature change, intensifying
cognition	thinking, judging, analyzing, doubting
communica-tion	telling, asking, ordering, singing
competition	fighting, athletic activities
consumption	eating and drinking
contact	touching, hitting, tying, digging
creation	sewing, baking, painting, performing
emotion	feeling
motion	walking, flying, swimming
perception	seeing, hearing, feeling
possession	buying, selling, owning
social	political and social activities and events
stative	being, having, spatial relations
weather	raining, snowing, thawing, thundering

General dictionaries

- Function words (e.g. *the*, *for*, *rather*, *than*)
  - Also called stopwords (often removed)
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - 2300 words
  - 70 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative", etc.
- NRC Emotion Lexicon[1]
  - 10,000 words coded along four emotional dimensions: joy–sadness, anger-fear, trust-disgust, anticipation-surprise
- Norms of valence, arousal, and dominance[2]
  - Code 14,000 words along three emotional dimensions: valence, arousal, dominance

Sentiment analysis is a very common use case for dictionaries.

- Extract a "tone" measure — positive, negative, or neutral.
- Let (w_i, s_i) be dictionary words and their sentiment scores $s_i \in [-1, 1]$.

  e.g., ("perfect", 0.8), ("awful", -0.9)
- For a phrase j, compute sentiment by averaging only over words found in the dictionary:

  $$s_j = \frac{1}{K_j} \sum_{i \in \mathcal{D}(j)} s_i,$$

  where $\mathcal{D}(j)$ are dictionary matches and K_j is the number of matches.
- Words not in the dictionary are skipped (or they contribute 0).

Application — What drives the radicalization of online protests?

- On Facebook, the Yellow Vests discussions became increasingly negative.
- Boyer et al. 2024 decompose this trend into two margins:
  - Extensive margin: active users become more radical on average.
  - Intensive margin: a given user becomes more likely to post radical messages over time.
- We estimate:

$$Y_{s,i,t} = \delta_i + \gamma_t + \varepsilon_s,$$

  where $Y_{s,i,t}$ is the sentence's sentiment, δ_i is a user fixed effect, and γ_t is a month fixed effect.
- Implied decomposition of average radicalism in month t:

$$\mathbb{E}_t[Y] = \mathbb{E}_t[\delta] + \gamma_t, \qquad \mathbb{E}_t[\delta] = \sum_i s_{i,t}\, \delta_i,$$

Examples of the most positive sentences:

honneur gilet jaune

mdr

bravo

mercii jeune meilleur facon aider progres meilleur monde

bravo gabin media honnete souhaite reussite merite equipe bravo gj

Examples of the most negative sentences:

macron demission

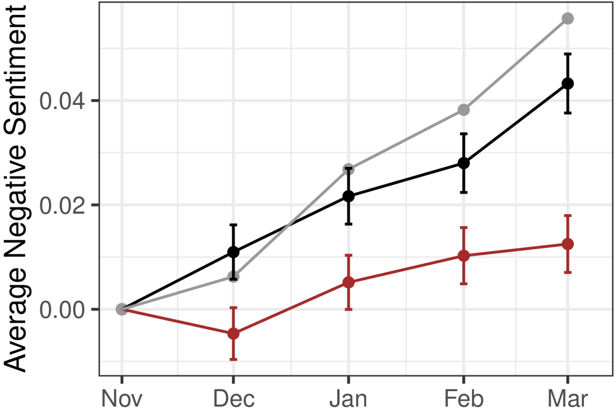
macron cabanon castananer enfer

florence menteur

bande pourriture batard

castaner assassin degage voleur menteur

Figure: Moderate users left, and those who remained radicalized.



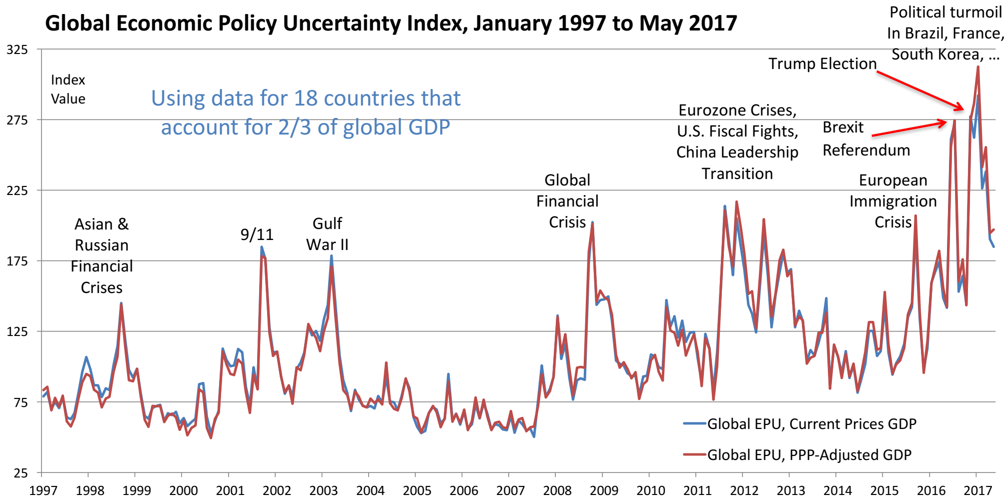
Notes: The red line is the composition effect. The black line is individual-level radicalization effect. The grey line is the total observed trend in the data.

Application — Measuring Economic Policy Uncertainty[3]

- Source: monthly token counts from 10 large U.S. newspapers.
- Step 1 (search rule): tag an article as EPU if it contains one uncertainty term (uncertainty/uncertain), one economy term (economic/economy), and one policy term (e.g., congress, deficit, Federal Reserve, legislation, regulation, White House).
- Step 2 (within-newspaper scaling): for each newspaper p and month t,

$$s_{p,t} = \frac{\#\text{EPU-tagged articles}_{p,t}}{\#\text{all articles}_{p,t}}.$$

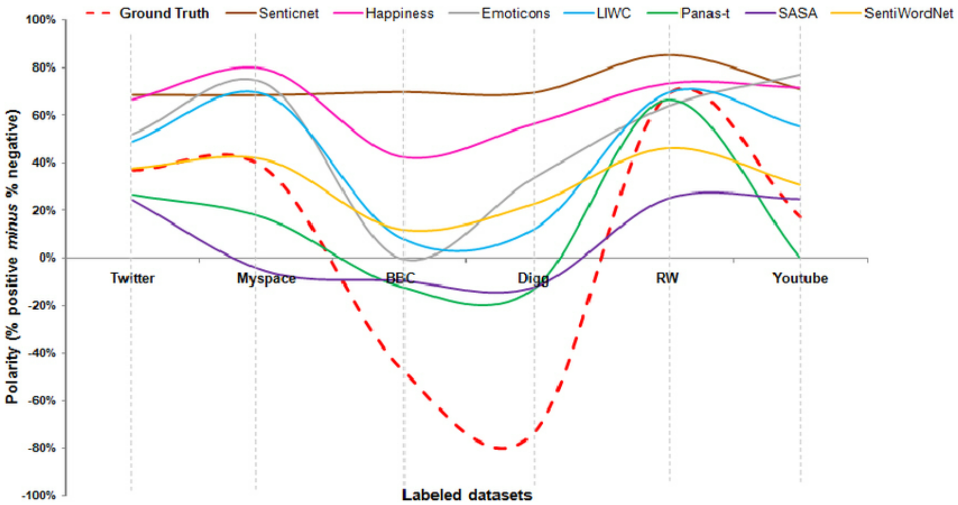
- Step 3 (aggregation and normalization): standardize each newspaper series to be comparable, average across newspapers, then rescale the final series to have mean 100 in a baseline period.



Notes: Global EPU calculated as the GDP-weighted average of monthly EPU index values for US, Canada, Brazil, Chile, UK, Germany, Italy, Spain, France, Netherlands, Russia, India, China, South Korea, Japan, Ireland, Sweden, and Australia, using GDP data from the IMF's World Economic Outlook Database. National EPU index values are from www.PolicyUncertainty.com and Baker, Bloom and Davis (2016). Each national EPU Index is renormalized to a mean of 100 from 1997 to 2015 before calculating the Global EPU Index.

Pros and Cons

- **Pros**
  - Straightforward and transparent
  - A lot of researcher control over the dictionary
- **Cons**
  - Requires domain-specific knowledge
  - Dictionaries cannot be exported easily to different contexts.
  - Predicted sentiment is sensitive to the choice of the dictionary.
  - Fails to identify irony.
  - No machine learning involved, so the model has no opportunity to discover patterns on its own.



Source/Notes: Polarity of the eight sentiment methods across the labeled datasets, indicating that existing methods vary widely in their agreement.[4]

Other Simple Metrics You Should Know

- Document length
- Word length
- Entropy: a measure of how evenly word usage is spread across the vocabulary. If $p(w)$ is the share of tokens that are word w in a document, then

  $$H = -\sum_{w \in V} p(w) \log p(w).$$

  Low entropy: repetition of a few words. High entropy: more diverse, evenly distributed word use.

References I

- Baker, Scott R, Nicholas Bloom, and Steven J Davis (2016). "Measuring economic policy uncertainty". In: *The quarterly journal of economics* 131.4, pp. 1593–1636.
- Boyer, Pierre C et al. (2024). *The lifecycle of protests in the digital age*. Tech. rep. CESifo Working Paper.
- Gonçalves, Pollyanna et al. (2013). "Comparing and combining sentiment analysis methods". In: *Proceedings of the first ACM conference on Online social networks*. ACM, pp. 27–38. DOI: 10.1145/2512938.2512951.
- Mohammad, Saif M and Peter D Turney (2013). "Nrc emotion lexicon". In: *National Research Council, Canada* 2, p. 234.
- Warriner, Amy Beth, Victor Kuperman, and Marc Brysbaert (2013). "Norms of valence, arousal, and dominance for 13,915 English lemmas". In: *Behavior research methods* 45.4, pp. 1191–1207.