Keyword Methods

HSS 510: NLP for HSS

Taegyoon Kim

Mar 20, 2024

Agenda

Things to be covered

- Principles of measurement
- Counting keywords
- Emotion/sentiment dictionary
- Moral foundation dictionary
- Validations and limitations
- Discriminating words (Fightin' Words)
- Guided coding: sentiment analysis (Python) & discriminating words (R)

Principles of measurement

Concepts and measures

- Conceptualization
  - E.g., “anti-vaccine rhetoric is speech that disputes the safety/efficacy of vaccines”
- Operationalization
  - E.g., identifying expressions that link autism to vaccination
- Measurement
  - E.g., assigning 1 to statements that link autism to vaccination

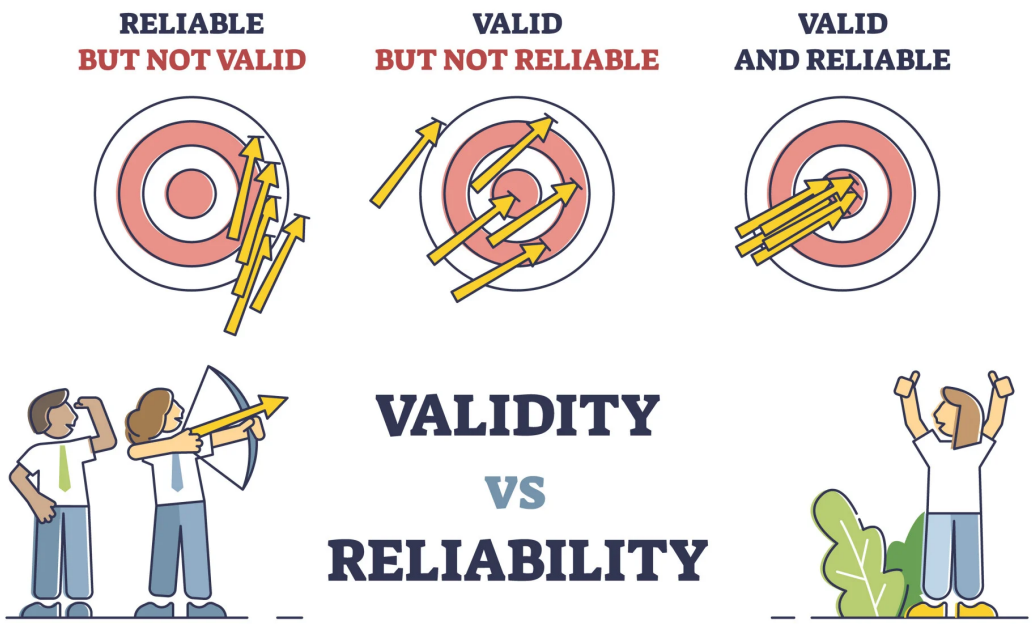
Principles of measurement

Many NLP models are used for measurement, including

- Text classification
- Topic models
- Document similarities (e.g., cosine similarity)
- Etc.

Principles of measurement

Reliability and validity



Keyword-based Methods

Supervised learning assigns texts into known categories

- Construct (or outsource) training data (documents are annotated)
- Build a statistical model that provides a mapping between texts and labels
- Label unseen texts (predict their labels)

Keyword-based methods offer alternatives

- Word-level approach
- A list/dictionary of keywords (sometimes with weights)
- Quick, clear, and easy to communicate (but usually less accurate)

Counting Keywords

E.g., we want to measure attention to (or interest in) presidential politics in U.S. news media

- Count the number of news articles that mention the words `president` or `White House`
- How can this approach go wrong?

Counting Keywords

False positive (Type I Error)

- Occurs when a test incorrectly indicates the presence of a condition when it is not actually present
  - E.g., Diagnose a patient has a disease when the patient does not

False negative (Type II Error)

- Occurs when a test fails to detect the presence of a condition when it is actually present
  - E.g., Diagnose a patient does not have a disease when the patient does

Couting Keywords

However, keywords carefully curated (along with validation) can be highly effective

- Validation: the process of assessing the degree of validity

Couting Keywords

Example I: Yian et al. (2021)

- Goal is to examine the use of scientific knowledge in policy during the COVID-19 pandemic
- Identifying policy documents and scientific publications about the COVID-19
  - Policy documents about the pandemic in many countries (**Overton**)
  - Scientific publications cited in the policy documents (**Dimensions**)

Couting Keywords

Example I: Yian et al. (2021)

- Keywords to identify policy documents about the pandemic (see here for non-English keywords)
  - `"2019-nCoV"`
  - `"COVID-19"`
  - `"coronavirus"`
  - `"corona virus"`
  - `“SARS-CoV-2”`

Couting Keywords

Example I: Yian et al. (2021)

- Keywords to identify scientific publications about the pandemic
  - `"2019-nCoV"`
  - `"COVID-19"`
  - `“SARS-CoV-2”`
  - `"HCoV-2019"`
  - `"hcov"`
  - `"NCOVID-19"`
  - `"severe acute respiratory syndrome coronavirus 2"`
  - `"severe acute respiratory syndrome corona virus 2"`
  - (`"coronavirus"` OR `"corona virus"`) AND (`Wuhan` OR `China` OR `novel`)

Couting Keywords

Note that humans are limited in recalling relevant keywords

- Relevant when you want to go as representative/comprehensive as possible
- This might lead to low recall (omission of relevant documents)
- Statistical approaches to deal with incomprehensiveness of human-generated keywords (e.g., King et al. (2017))

Couting keywords

Example II: Duneier (2016)

- Traces historical contexts in which the word `"ghetto"` is used
  - Frequently used to refer to the Nazi's ghettoization of the Jews
  - Subsequently also used to refer to the segregation of the Black in U.S.

Couting Keywords

Example II: Duneier (2016)

- According to Duneier, the former *resulted in* the latter

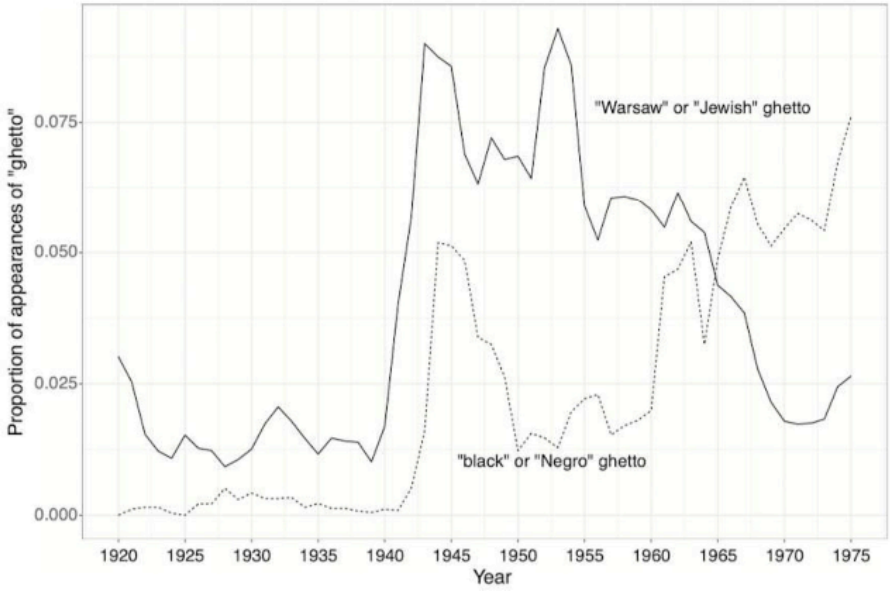


Figure 16.1. Graph 1 from Duneier (2016) depicting the proportion of the uses of the word ghetto in context of words signalling one of two uses. Data derived from Google Ngrams.

Couting Keywords

Example II: Duneier (2016)

- Duneier conducted close reading of documents that contribute to the surge
  - E.g., Cayton and Drake (1945): invokes the concept of the Black Ghetto in order to draw attention to the coercive housing policies to segregate Black people in U.S.
  - The authors drew on the association of the word `"ghetto"` with the Nazi oppression of the Jews to highlight the hypocrisy of America's treatment
- Importantly, the Black segregation started as early as 1928

Couting Keywords

Example II: Duneier (2016)

- Demonstrates the power of simple keyword counting combined with qualitative evidence

  → Counts of four phrases + deep reading on the underlying sources

Dictionary Methods

Generalization of keyword counting

- The previous examples of the use of keywords are often *ad hoc*
- Pre-defined set of keywords associated with certain concepts like sentiment
- We can measure *a host of* concepts using dictionary methods
  - Sentiment, emotion, morality, aggression, etc.
- The goal of dictionary methods is often identical to supervised learning
  - Assign documents into known categories
- High level of domain expertise / qualitative judgment can be required

Dictionary Methods

Two approaches

- We can construct our own dictionary
- Alternatively (and much more common), there are a variety of off-the-shelf dictionaries

Dictionary Methods

E.g., measuring positive sentiment in newspaper articles

- We find a dictionary of keywords with each associated with a score signaling the degree of pleasantness (e.g., “fun”, “excellent”, etc.)
- Count the number of times each keywords appears and add them up
- (Optional) we normalize by document length (word count)

Dictionaries for Affect

Many concepts and meanings are named under *affect*

- *Affect* is a term that encompasses emotion, sentiment, personality, mood, attitudes, etc. (Picard, 1995; Scherer, 2000)
- See [JM] Chp. 25 (Section 25.1) for more theoretical discussions of emotions

Dictionaries for Affect

Keywords in affect dictionaries

- Affective lexicons: keywords that carry particularly strong cues to affect meanings (= connotations)

Linguistic Inquiry and Word Count (LIWC)

Most recent version (LIWC-22) (proprietary)

- Not just affect: + 100 linguistic dimensions
  - E.g., positive/negative, moralization, I-word, etc.
- Uses a dictionary to calculate the percentage of words in the text that match certain dimension

Linguistic Inquiry and Word Count (LIWC)

E.g., "I am very disappointed"

Your text sample is 4 words. The LIWC-22 analysis of the text sample you entered is below. Note that LIWC-22 actually produces about 100 different output dimensions. Remember: the more text that you have available for analysis, the more trustworthy and reliable your results will be.

RESULTS

Traditional LIWC Dimension	Your Text	Average for Social Media Language
I-words (I, me, my)	25.00	5.44
Positive Tone	0.00	5.93
Negative Tone	25.00	2.34
Social Words	0.00	6.74
Cognitive Processes	0.00	8.86
Allure	0.00	8.62
Moralization	0.00	0.27
Summary Variables		
Analytic	0.00	47.06
Authentic	89.41	62.38

Traditional LIWC dimensions reflect percentage of total words within the text you provided. The Summary Variables are composites derived from scientific research that have been converted to 100-point scales, where 0 = "very low" along the dimension and 100 = "very high." Analytic refers to analytical or formal thinking. Authentic is a property of language that reflects when someone is speaking in an unfiltered, off-the-cuff fashion.

Want to learn more about the meaning of the LIWC output? See: Interpreting LIWC Output.

ANALYZE ANOTHER TEXT

Linguistic Inquiry and Word Count (LIWC)

Very widely used for tasks including

- Detecting political sentiment from tweets (Tumasjan et al. 2010)
- Predicting the onset of depression in individuals based on text from social media (De Choudhury et al. 2013)
- Differentiating happy romantic couples from unhappy ones based on their instant messages (Hancock et al. 2007)

Linguistic Inquiry and Word Count (LIWC)

Kramer et al. (2014)

- Examines emotional contagion on Facebook
- N = 689, 003 Facebook users
- Manipulated content shown on news feeds to test emotional contagion hypothesis
- Treatment 1: positive content more visible on news feed
- Treatment 2: negative content more visible on news feed
- Control: no news feed intervention

Linguistic Inquiry and Word Count (LIWC)

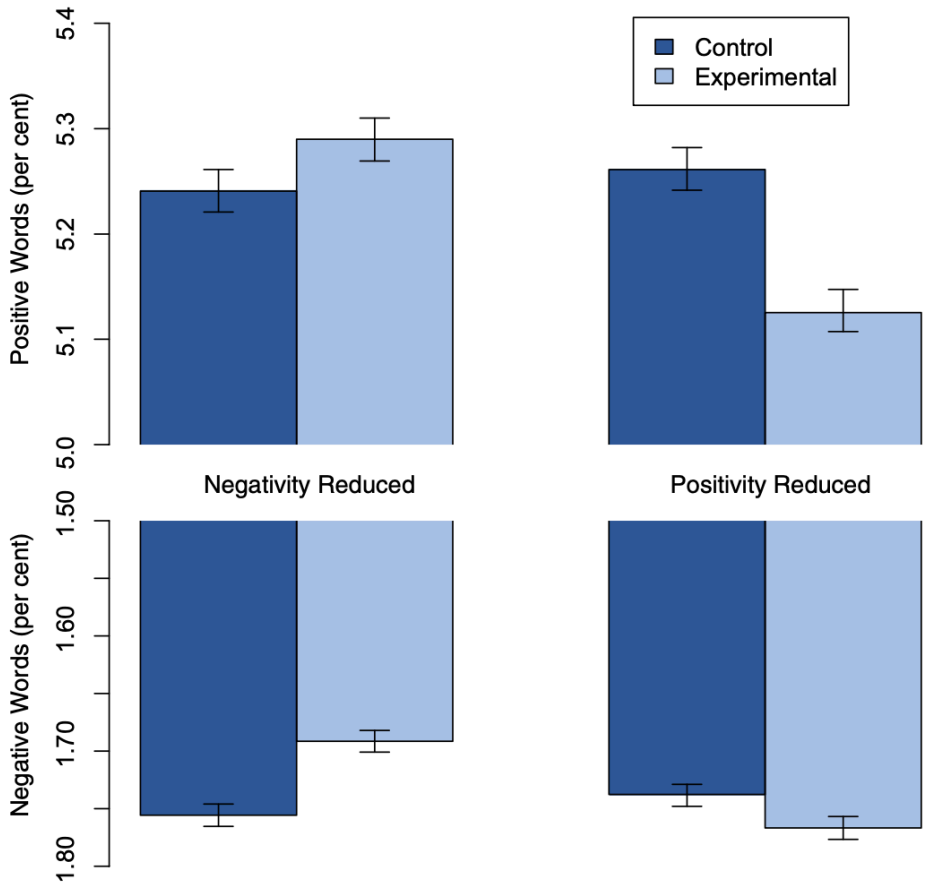
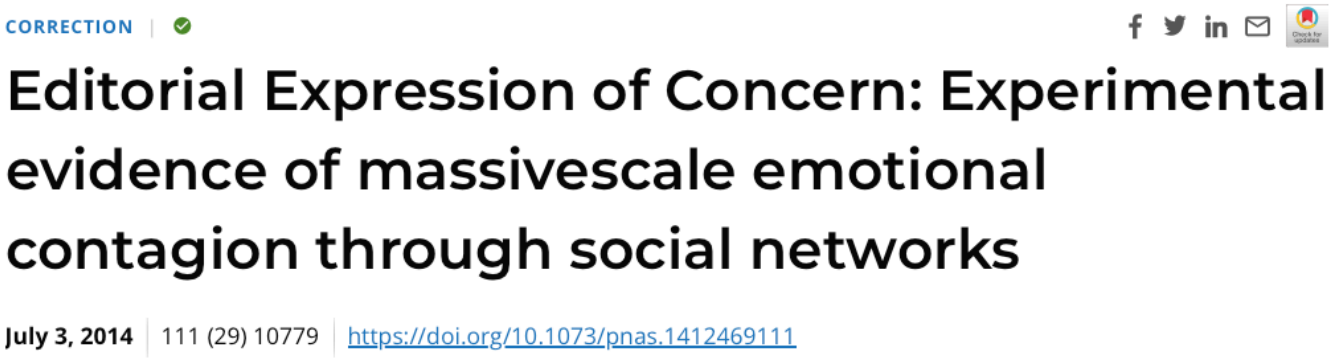


Fig. 1. Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

Linguistic Inquiry and Word Count (LIWC)

Huge concerns were raised about the ethics of the study



CORRECTION

Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks

July 3, 2014 | 111 (29) 10779 | https://doi.org/10.1073/pnas.1412469111

Valence Aware Dictionary and sEntiment Reasoner (VADER)

Hutto et al.(2014)

- Improves LIWC and other sentiment dictionaries focused on social media
- Not just polarity but also intensity
- Initialisms, emoticons, or slangs
- Crowd-sourced labeling of keywords

Valence Aware Dictionary and sEntiment Reasoner (VADER)

Hutto et al.(2014)

- Grammatical and syntactical cues
  - Punctuation: good vs. good!!!
  - Capitalization: great vs. GREAT
  - Degree modifiers: extremely good vs. good
  - Contrastive conjunction: “The food here is great, but the service is horrible”
  - Negation: “The food here isn’t really all that great”

Valence Aware Dictionary and sEntiment Reasoner (VADER)

Example keywords and weights

- `great`: 3.1
- `good`: 1.9
- `sucks`/`sux`: –1.5
- `:(`: –2.2
- `horrible`: –2.5

Valence Aware Dictionary and sEntiment Reasoner (VADER)

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 4: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).

Valence Aware Dictionary and sEntiment Reasoner (VADER)

	3-Class Classification Accuracy (F1 scores) Test Sets			
	Tweets	Movie	Amazon	NYT
VADER	**0.96**	0.61	**0.63**	**0.55**
NB (tweets)	0.84	0.53	0.53	0.42
ME (tweets)	0.83	0.56	0.58	0.45
SVM-C (tweets)	0.83	0.56	0.55	0.46
SVM-R (tweets)	0.65	0.49	0.51	0.46
NB (movie)	0.56	**0.75**	0.49	0.44
ME (movie)	0.56	**0.75**	0.51	0.45
NB (amazon)	0.69	0.55	0.61	0.48
ME (amazon)	0.67	0.55	0.60	0.43
SVM-C (amazon)	0.64	0.55	0.58	0.42
SVM-R (amazon)	0.54	0.49	0.48	0.44
NB (nyt)	0.59	0.56	0.51	0.49
ME (nyt)	0.58	0.55	0.51	0.50

Table 5: Three-class accuracy (F1 scores) for each machine trained model (and the corpus it was trained on) as tested against every other domain context (SVM models for the movie and NYT data were too intensive for our multicore CPUs with 94GB RAM)

Dictionaries for Measuring Moral Rhetoric

Moral Foundations Theory (MFT)

- Morality: principles concerning the distinction between right and wrong behavior
- MFT: (Social psychological) theory designed to understand the origins of human morality

Dictionaries for Measuring Moral Rhetoric

Moral Foundations Theory (MFT)

- There are five innate, universal psychological systems that underlie morality
  - Care/harm (the desire to protect individuals from harm)
  - Fairness/cheating (the desire for justice and reciprocity)
  - Loyalty/betrayal (the value of loyalty and group cohesion)
  - Authority/subversion (the value of respect for tradition and authority)
  - Sanctity/degradation (the value of purity and avoidance of contamination)

Dictionaries for Measuring Moral Rhetoric

MFD (**M**oral **F**oundations **D**ictionary) (Frimer et al., 2019)

- Version 1.0: link
- Version 2.0: link

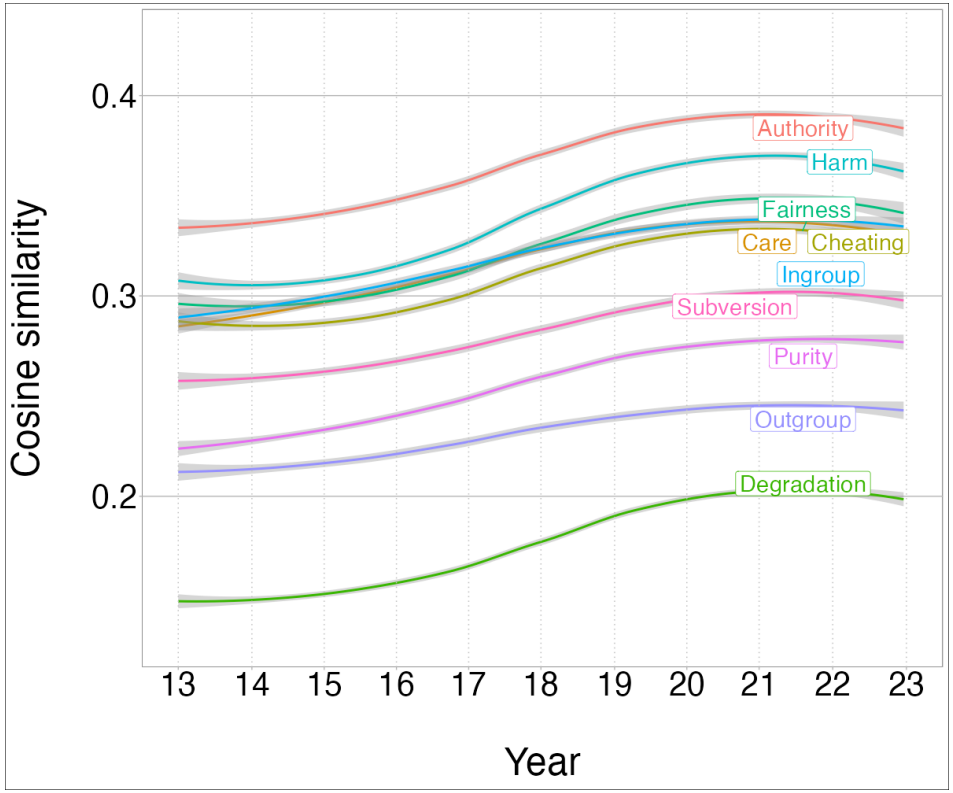
Dictionaries for Measuring Moral Rhetoric

More recent approaches to measuring morality

- Dictionaries combined with word embeddings
  - E.g., similarity between
    - The centroid of embeddings for terms in MFD, on one hand
    - The centroid of the embeddings of the terms in a target document, on the other hand
- Human annotations and machine learning (including LLMs)
  - E.g., recent work based on variants of GPT: Rathje et al. (2024)

Dictionaries for Measuring Moral Rhetoric

US legislators' moral rhetoric (Twitter, dictionary + embedding)



Validating Dictionaries

- Validation
  - (Semi-)randomly sample a small number of texts
  - Generate human labels
  - Compare dictionary-based labels and human labels
    - We will discuss useful metrics next week
- Note that dictionaries do not always travel across contexts
  - E.g., “tax”, “cost”, “cancer” are considered signaling negative connotations, but they do not in accounting/finance
- Easier when the dictionary and the research objectives are highly aligned
  - E.g., keywords related to the COVID-19 from Yian et al. (2021)

Limitations of Dictionaries

Barbera et al. (2021): sentiment in NYT articles

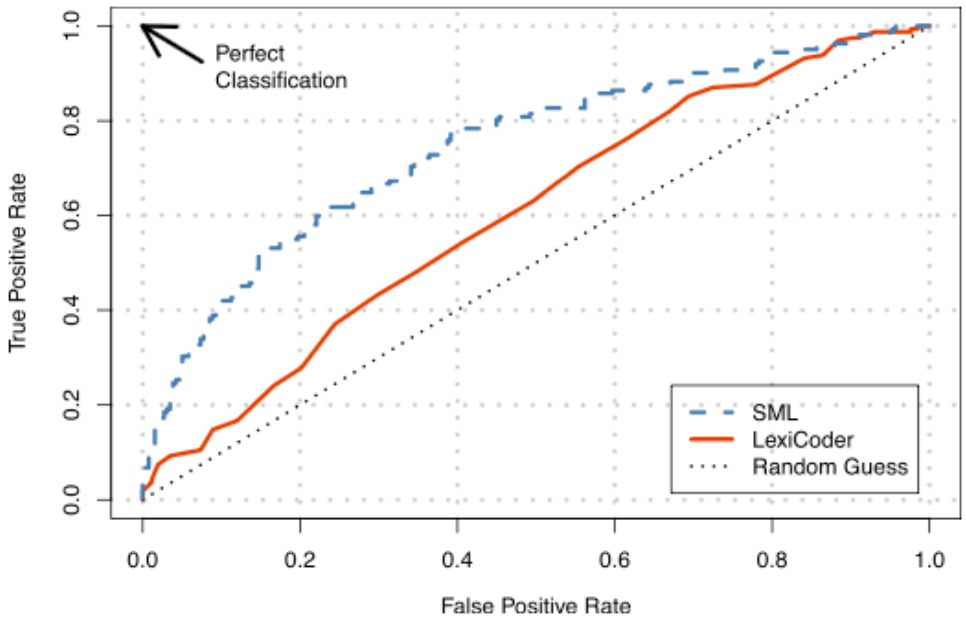


Figure 5. Receiver Operator Characteristic Curve: Lexicoder versus SML.
Note: The x-axis gives the false positive rate—the proportion of all negatively toned articles in CF Truth that were classified as positively toned—and the y-axis gives the true positive rate—the proportion of all positively toned articles in CF Truth that were classified as positive. Each point on the curve represents the misclassification rate for a given classification threshold. The corpus used in the analysis is based on the keyword search of *The New York Times* 1980–2011.

Limitations of Dictionaries

Widmann and Wich (2023): classifying emotions

Emotions	Actual	Predicted	Precision	Recall	F1
		ed8 dictionary			
Anger	508	281	0.83	0.46	0.59
Fear	189	287	0.43	0.66	0.52
Disgust	86	182	0.30	0.63	0.40
Sadness	201	289	0.41	0.59	0.48
Joy	143	179	0.46	0.58	0.52
Enthusiasm	220	248	0.44	0.50	0.47
Pride	158	247	0.31	0.48	0.38
Hope	305	303	0.53	0.53	0.53
		Word-embeddings-based neural network approach			
Anger	508	500	0.80	0.78	0.79
Fear	189	152	0.61	0.49	0.55
Disgust	86	67	0.60	0.47	0.52
Sadness	201	122	0.70	0.42	0.53
Joy	143	92	0.68	0.44	0.54
Enthusiasm	220	176	0.64	0.51	0.57
Pride	158	123	0.52	0.41	0.46
Hope	305	265	0.69	0.60	0.64
		Transformer-based (ELECTRA) approach			
Anger	508	495	0.85	0.83	0.84
Fear	189	221	0.60	0.70	0.64
Disgust	86	89	0.61	0.63	0.62
Sadness	201	181	0.64	0.57	0.60
Joy	143	122	0.70	0.59	0.64
Enthusiasm	220	242	0.62	0.68	0.65
Pride	158	151	0.61	0.58	0.60
Hope	305	352	0.68	0.78	0.73

Limitations of Dictionaries

Dictionaries can still be useful

- Low resource (no need to build a training set)
- Fast implementation and less computing power
- Transparent (much less black-boxy)
- Useful for preliminary analysis

Discriminating Words

How can we identify words that distinguish two groups?

- Political parties
- Gender identities
- Periods or generations

Discriminating Words

The resulting keywords are of interest in and of itself

- Serve as a lens through which to examine framing, topic, etc.
- E.g., how do Democrats and Republicans in the U.S. talk differently about abortion?
- We can apply this to generate keywords, which in turn can be used as a dictionary (p. 181 in [**GRS**])
  - E.g., pro-vaccine vs. anti-vaccine comments during the COVID-19 pandemic

Discriminating Words

Some approaches

- Difference of frequencies: $|f^i(horrible) - f^j(horrible)|$
- Difference in proportions: $|p^i(horrible) - p^j(horrible)|$
- Classification: take weights from “g = f(w)”

→ See Monroe et al. (2008) for detailed discussions

Discriminating Words

Fightin' Words (Monroe et al., 2008)

- Words that are used differently by two political parties
- The extent to which each word is used differently by two political parties

Discriminating Words

The log odds ratio (LOR) of the word “horrible” is given by:

$$LOR(horrible) = \log\left(\frac{p^i(horrible)}{1 - p^i(horrible)}\right) - \log\left(\frac{p^j(horrible)}{1 - p^j(horrible)}\right)$$

Which simplifies to:

$$= \log\left(\frac{f^i(horrible)}{n^i - f^i(horrible)}\right) - \log\left(\frac{f^j(horrible)}{n^j - f^j(horrible)}\right)$$

Where $p^i(horrible)$ and $p^j(horrible)$ are probabilities of “horrible” in corpus **i** and **j**, f^i and n^j are the number of times “horrible” appears in the respective corpus, and n^i and n^j are the numbers of words in the respective corpus.

To further incorporate a prior estimate of what we expect the frequency of each word **w** to be, we add the counts from the entire corpus to the numerator and denominator:

$$\delta_w^{(i-j)} = \log\left(\frac{f_w^i + \alpha_w}{n_i + \alpha_0 - (f_w^i + \alpha_w)}\right) - \log\left(\frac{f_w^j + \alpha_w}{n_j + \alpha_0 - (f_w^j + \alpha_w)}\right)$$

We also need an estimate for the variance of the log-odds-ratio:

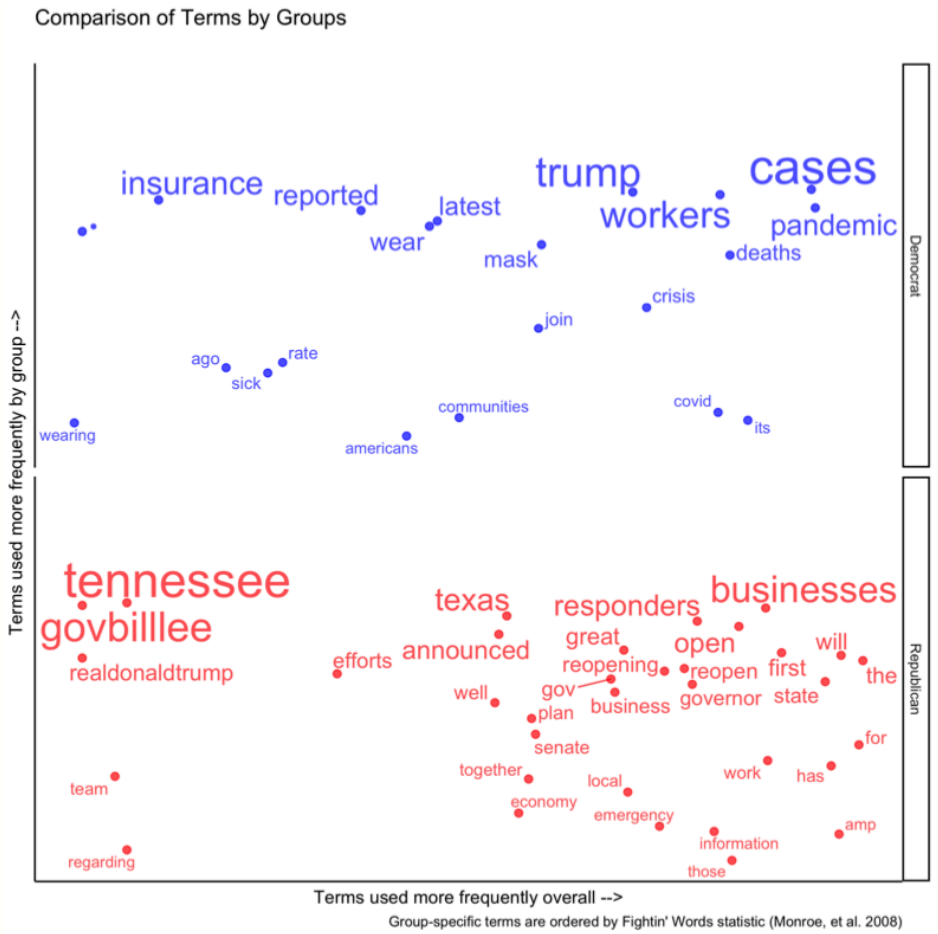
$$\sigma^2(\hat{\delta_w}^{(i-j)}) \approx \frac{1}{f_w^i + \alpha_w} + \frac{1}{f_w^j + \alpha_w}$$

The final statistic of interest (zeta or z-score):

$$\hat{\zeta_w}^{(i-j)} = \frac{\hat{\delta_w}^{(i-j)}}{\sqrt{\sigma^2(\hat{\delta_w}^{(i-j)})}}$$

Discriminating Words

Tweets about the pandemic: Democrats vs. Republicans



Summary

Counting theoretically-relevant keywords can be highly effective

Dictionaries are increasingly less useful but still serve useful roles

Keywords that discriminate between groups can be useful

Guide coding

VADER in Python (link) and Fightin' Words in R (link)