Conspiracy Language Analysis

Sen Sub Laban LING 2340 December 8, 2022

Background

- Conspiracy theories (CTs) are narratives that attempt to explain significant social events as being secretly plotted by powerful and malicious elites at the expense of an unwitting population.
- Belief in CTs is widespread, with over 50% of the US population believing in at least one in 2013.
- A prime example: In 2020, in the middle of the COVID-19 pandemic, health-related misinformation attracted four times as much traffic as official health sources on social media.
- Research has shown that belief in CTs is correlated to rejection of official information and science, detrimental or lack of social action, general distrust and political alienation, justification for engaging in crime, and anti-Semitic and Islamophobic attitudes.
- The internet is the prime vehicle through which conspiracy theories spread in the modern day. CT *beliefs* do not spread per se, but rather, CTs spread as materialized narratives, so research must investigate the *content of CT narratives* in order to better understand their spread.

Research questions

Which lexical features of documents that are considered conspiratorial have the strongest influence on social media engagement?

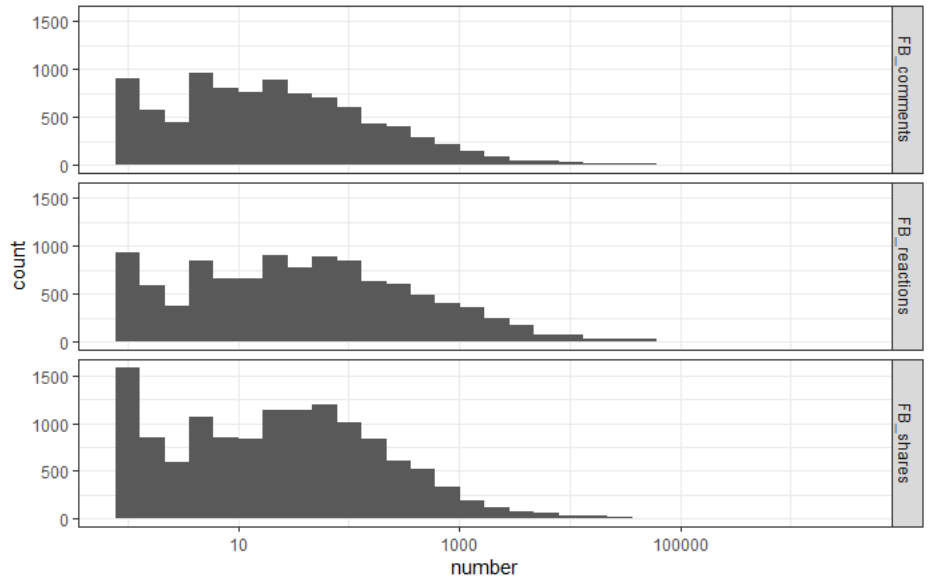
Does this differ from the LFs that influence social media engagement with documents that are non-conspiratorial? If so, how?

Language of Conspiracy Corpus (LOCO)

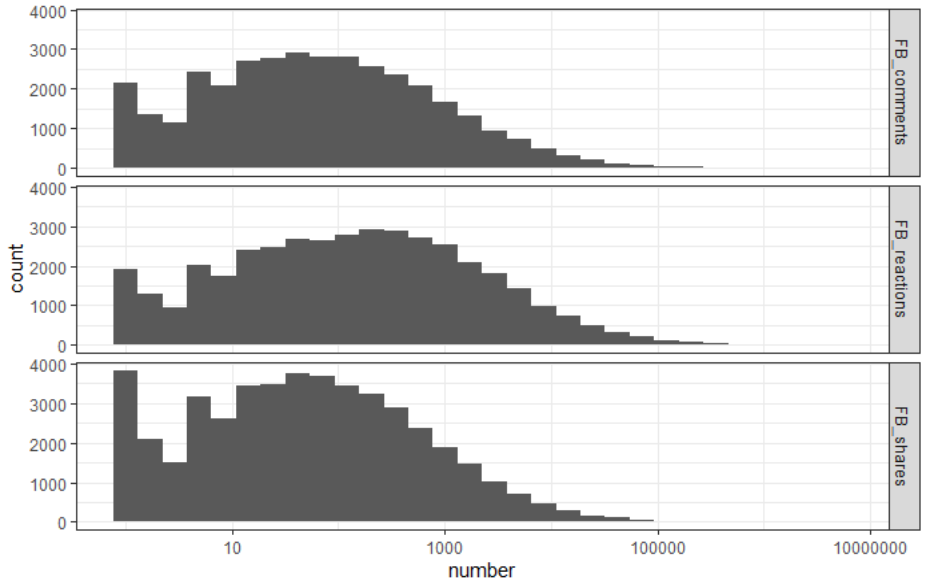
	Mainstream	Conspiracy	Whole corpus
No. of documents	72, 806	23, 937	96, 743
No. of websites	92	58	150
Range of years	1853—2020	2004—2020	1853—2020
	M (SD) [range]	*M (SD)* [range]	*M (SD)* [range]
Words per document	805.94 (939) [97-9507]	1236.32 (1307) [100-9428]	912.43 (1059) [97-9507]
Total no. of words	58, 677, 322	29, 593, 678	88, 271, 000

Social media engagement

Conspiracy



Mainstream



Lexical feature extraction

LIWC

- Linguistic Inquiry and Word Count
- Standardized output: no. of words in given category divided by the total no. of words from text file.
- Returns percentages (range: 0-100)
- Detects grammatical categories such as articles, prepositions, pronouns, etc.
- Built on human coding
- 93 categories

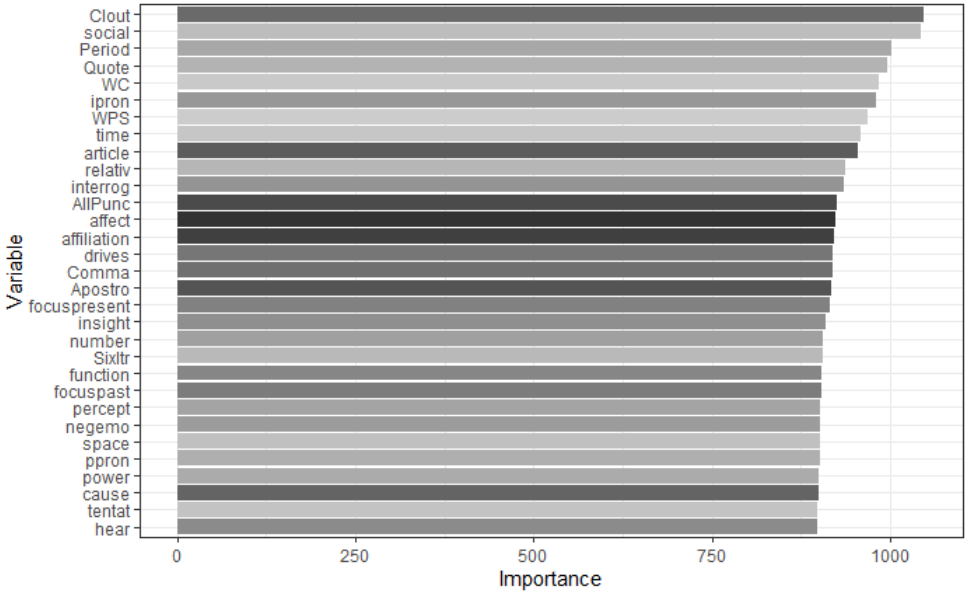
Empath

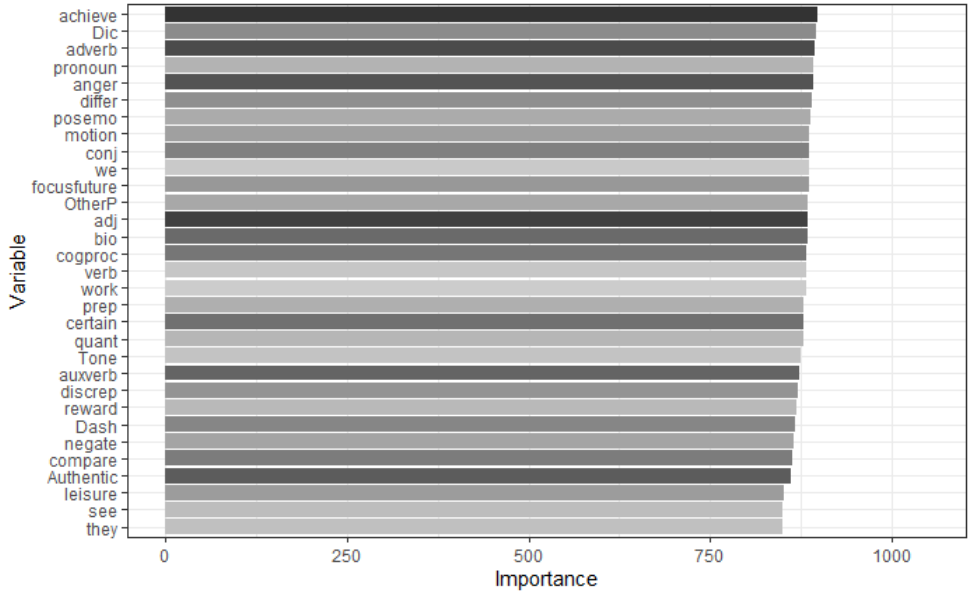
- Emotion-based Empath categories serve as something of a proxy for sentiment analysis.
- Standardized output: no. of words in given category divided by the total no. of words from text file.
- Returns ratios (range: 0-1)
- No grammatical categories
- Categories built in a data-driven fashion from a semantic database
- 194 categories

Random forest

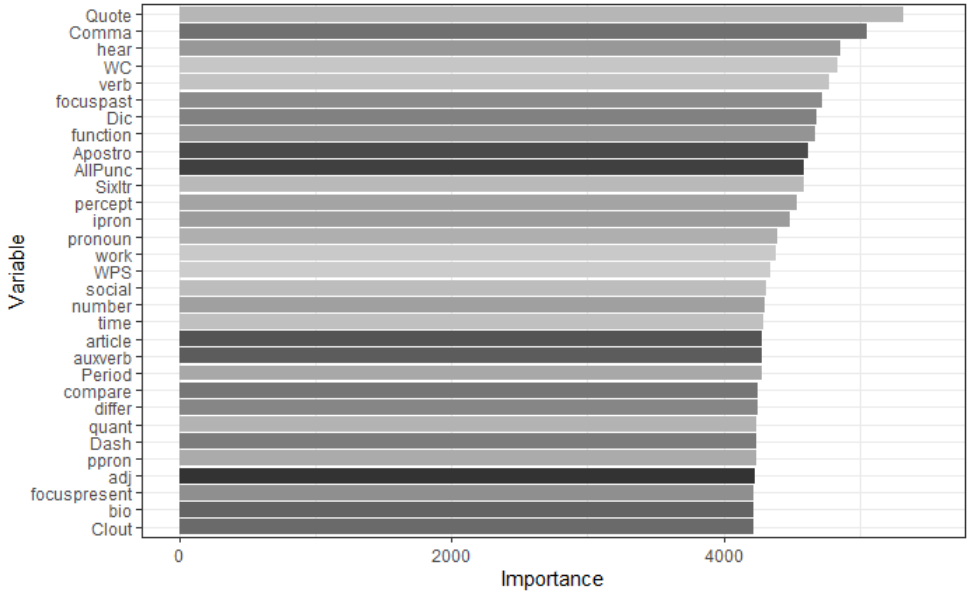
- Totaled the sum of the FB shares, comments, and reactions per document to create a new variable, engagements
- Took the natural log to account for the strong skew in engagement numbers
- Set a seed and generated a random forest to calculate the importance (predictive power) of all 93 LIWC variables on log engagement

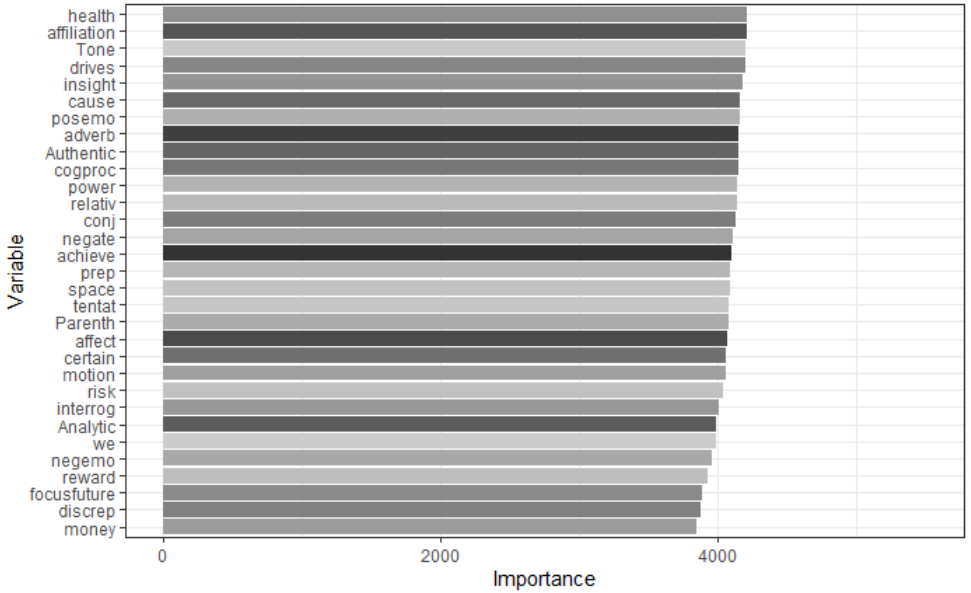
Variable importance (Conspiracy)





Variable importance (Mainstream)





Variable importance comparison

Conspiracy

Rank <int>	Variable <chr>	Importance <dbl>
1	Clout	1047.23192
2	social	1044.17450
3	Period	1002.73905
4	Quote	996.21255
5	WC	985.08438
6	ipron	980.74137
7	WPS	968.11126
8	time	959.13529
9	article	955.43096
10	relativ	937.88837

1-10 of 93 rows

Mainstream

Rank <int>	Variable <chr>	Importance <dbl>
1	Quote	5323.6610
2	Comma	5048.2962
3	hear	4858.6038
4	WC	4836.2173
5	verb	4770.5403
6	focuspast	4724.6656
7	Dic	4686.0054
8	function	4676.7556
9	Apostro	4615.9614
10	AllPunc	4591.4825

1-10 of 93 rows

Next steps

More statistical analysis

Select variables with high predictive power and fit a linear mixed-effects model

Account for variance inflation factor

Other statistical models?

Limitations

Lexical analyses using a bag-of-words approach omit important context and can overlook subtle differences in how topics are discussed

References and further reading

Klein, C., Clutton, P., & Dunn, A. (2019). Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PLOS ONE*. 14(11): e0225098. 10.1371/journal.pone.0225098

Miani, A., Hills, T., & Bangerter, A. (2022). LOCO: The 88-million-word language of conspiracy corpus. *Behavior research methods, 54*(4), 1794–1817. https://doi.org/10.3758/s13428-021-01698-z

Mompelat et al. (2022). How "Loco" is the LOCO corpus? Annotating the language of conspiracy theories. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, 111–119. European Language Resources Association

Wood, M. & Douglas, K. (2015). Online communication as a window to conspiracist worldviews. *Frontiers in Psychology*. 6. 10.3389/fpsyg.2015.00836.