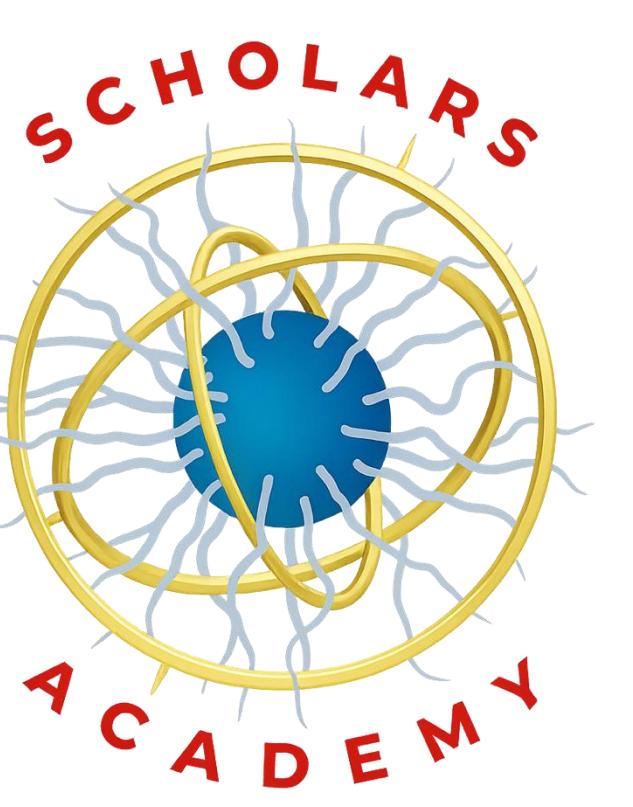


# Designing a reproducible NLP workflow for social media sentiment and topic modeling

Thomas Linden, Katherine Shoemaker, PhD

Department of Mathematics & Statistics, University of Houston – Downtown, Texas USA



## Background

- X (formerly Twitter) functions as a real-time reputation marketplace. For advertisers, brand/celebrity PR, and political offices, it acts like a nationwide focus group, surfacing what people notice, celebrate, or push back on within hours.
- But attention is fragmented across subcultures, humor and sarcasm mask intent, and coordinated activity can warp the apparent mood.
- Stakeholders need snapshots that are comparable over time and across audiences to understand where conversation is heading—not just how loud it is.

## Problem Statement

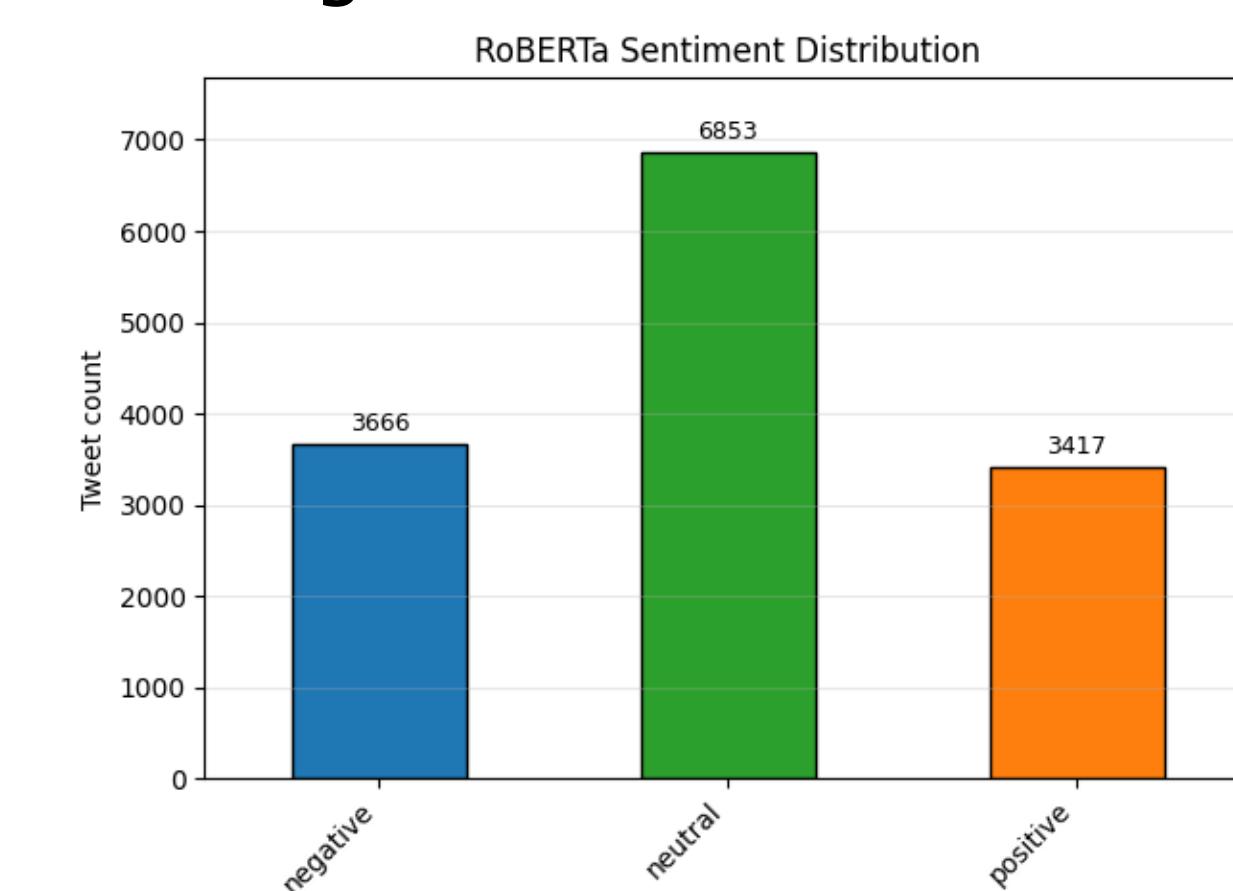
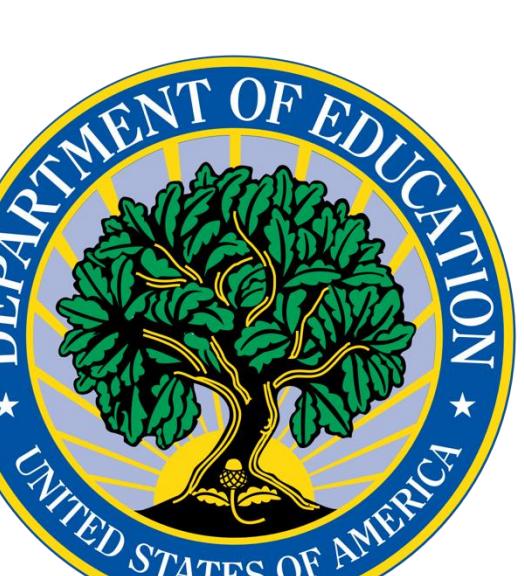
- We must determine when co-domain analysis (looking at what the *same users* discuss beyond the target keywords) adds decision value over a standard domain view (on-topic keywords only).
- Domain sampling gives crisp, on-message pulse checks; co-domain may reveal adjacent interests, early risks, and opportunity coalitions.
- The practical question: When does adding co-domain improve targeting, message tests, and crisis prep—and when is it just noise?

## Objectives

- Collect and preprocess X posts for analysis.
- Produce trustworthy pulse checks using transformer-based sentiment scoring.
- Extract what people talk about via TF-IDF metrics + LDA topic summaries.
- Compare positive vs. negative slices to spot drivers of praise vs. criticism.
- Output results for LLM synthesis.
- Compare domain vs. co-domain results.

## GitHub

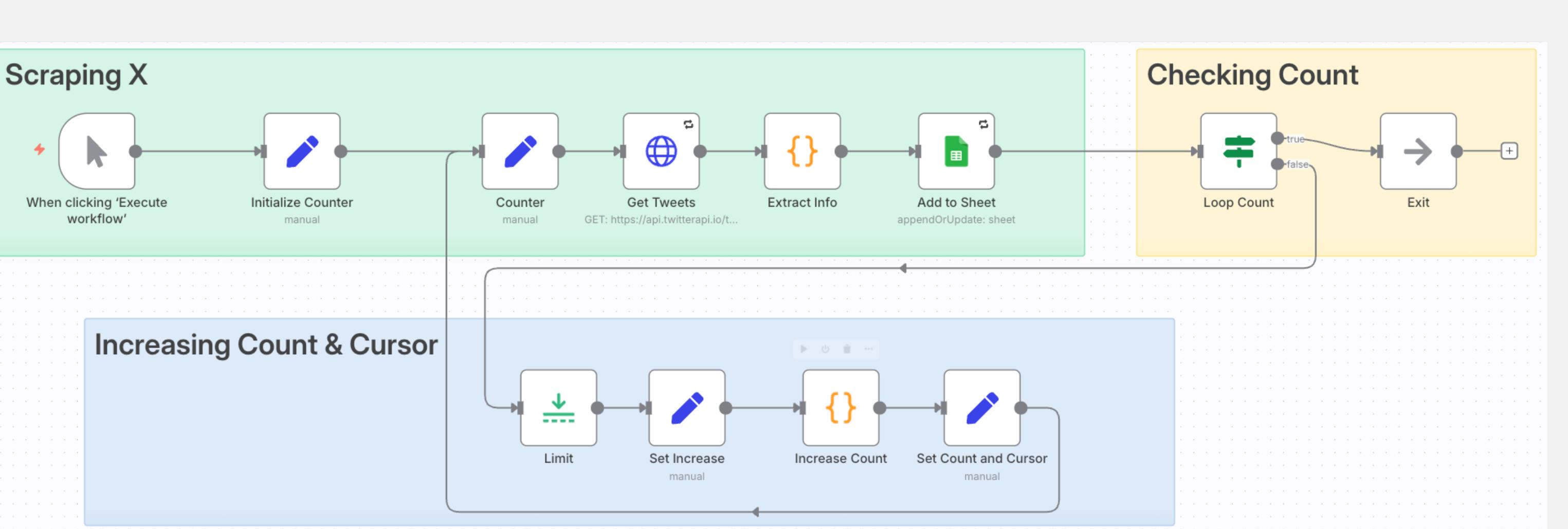
For co-domain and domain results of all 3 topics-  
<https://github.com/TLindenIII/Social-Media-Topic-Modeling>



## Methods

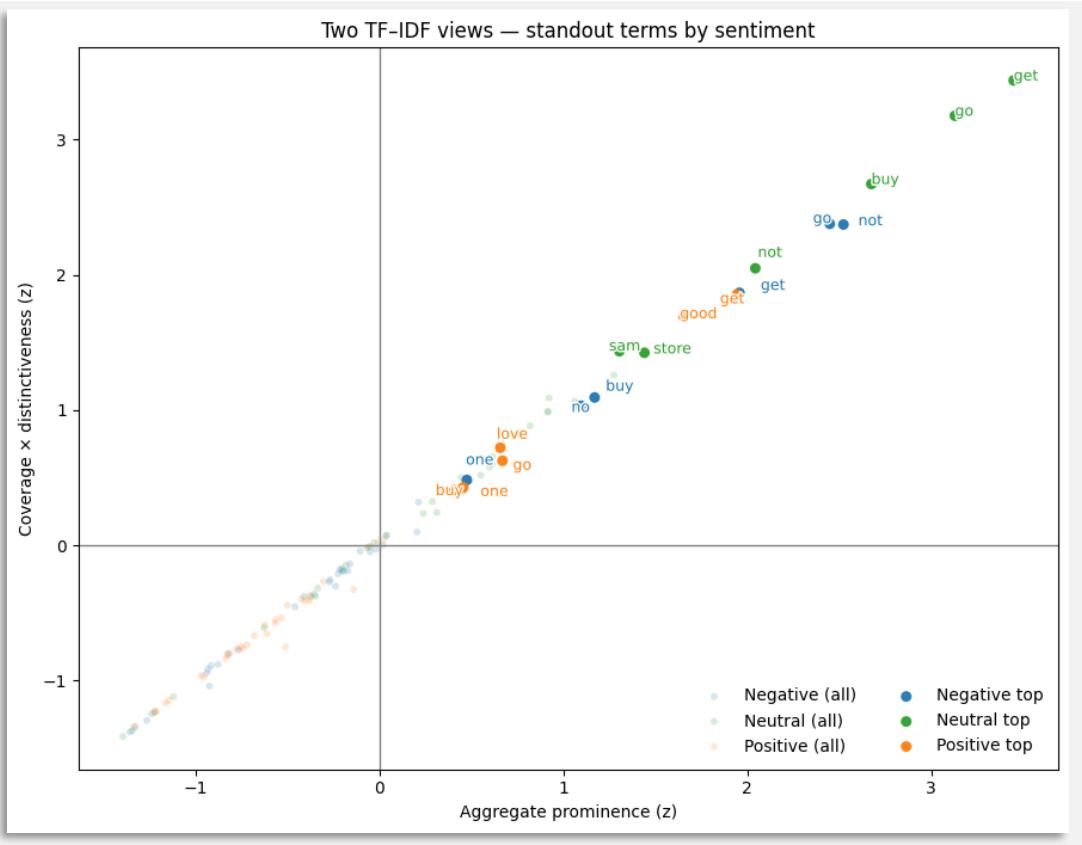
### Data

- twitterapi.io
- n8n: node-based automation
- 60k posts across NFL, Keir Starmer, & Costco “Domains”
- 5-day post window



### Key Terms

- Term Frequency-Inverse Document Frequency
- Compared Aggregate vs. weighted IDF: Spearman-rho of .999

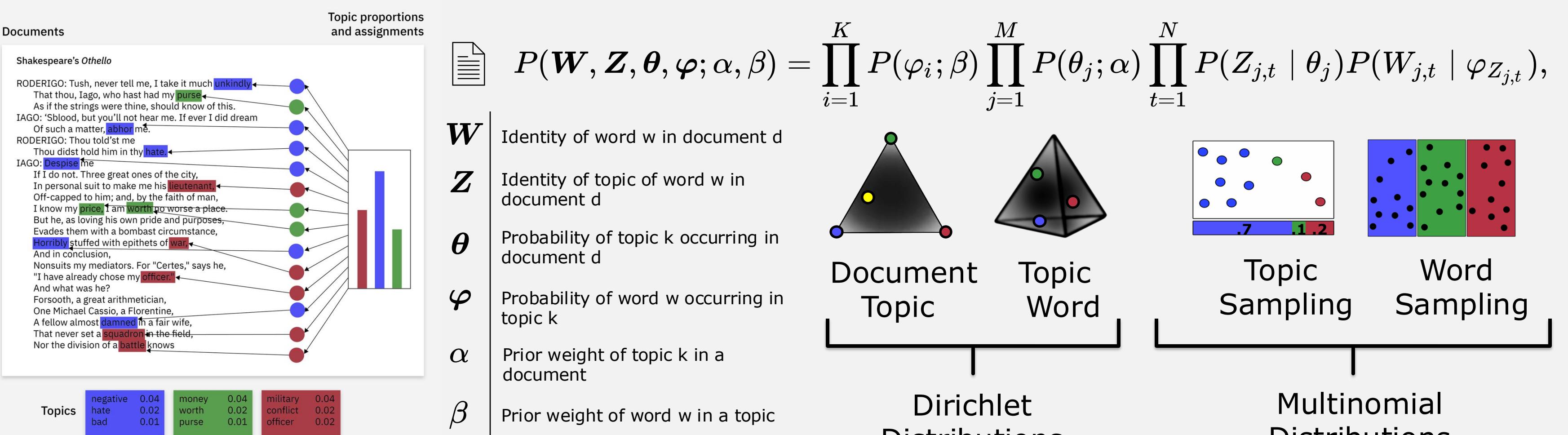


$$TFIDF_{t,d} = tf_{t,d} \times \log \left( \frac{N}{|\{d \in D : t \in d\}|} \right)$$

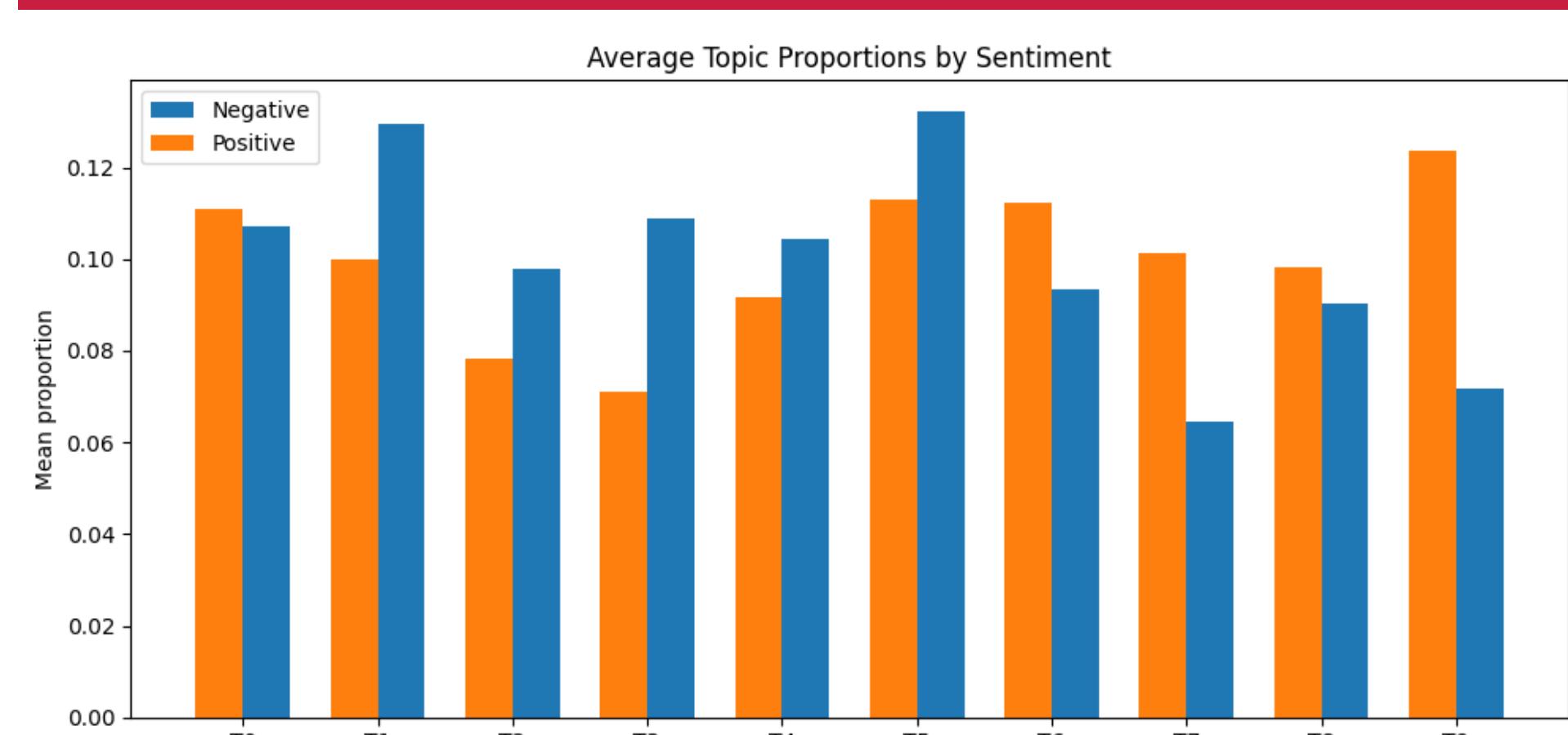
$t$  = term  
 $d$  = document  
 $tf$  = frequency of term  $t$  in document  $d$   
 $N$  = total number of documents  
 $D$  = number of documents  $d$  containing term  $t$

### Modeling

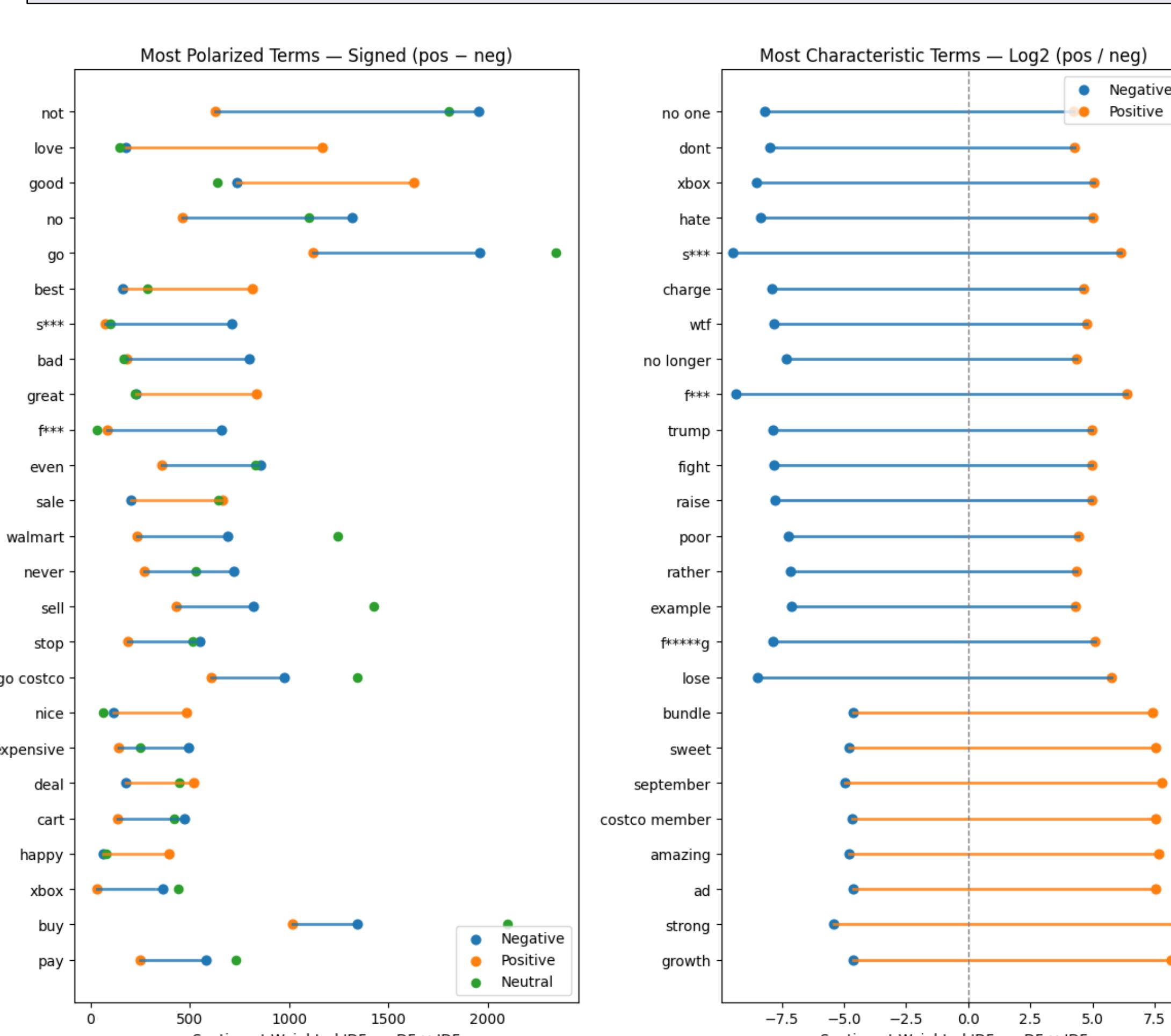
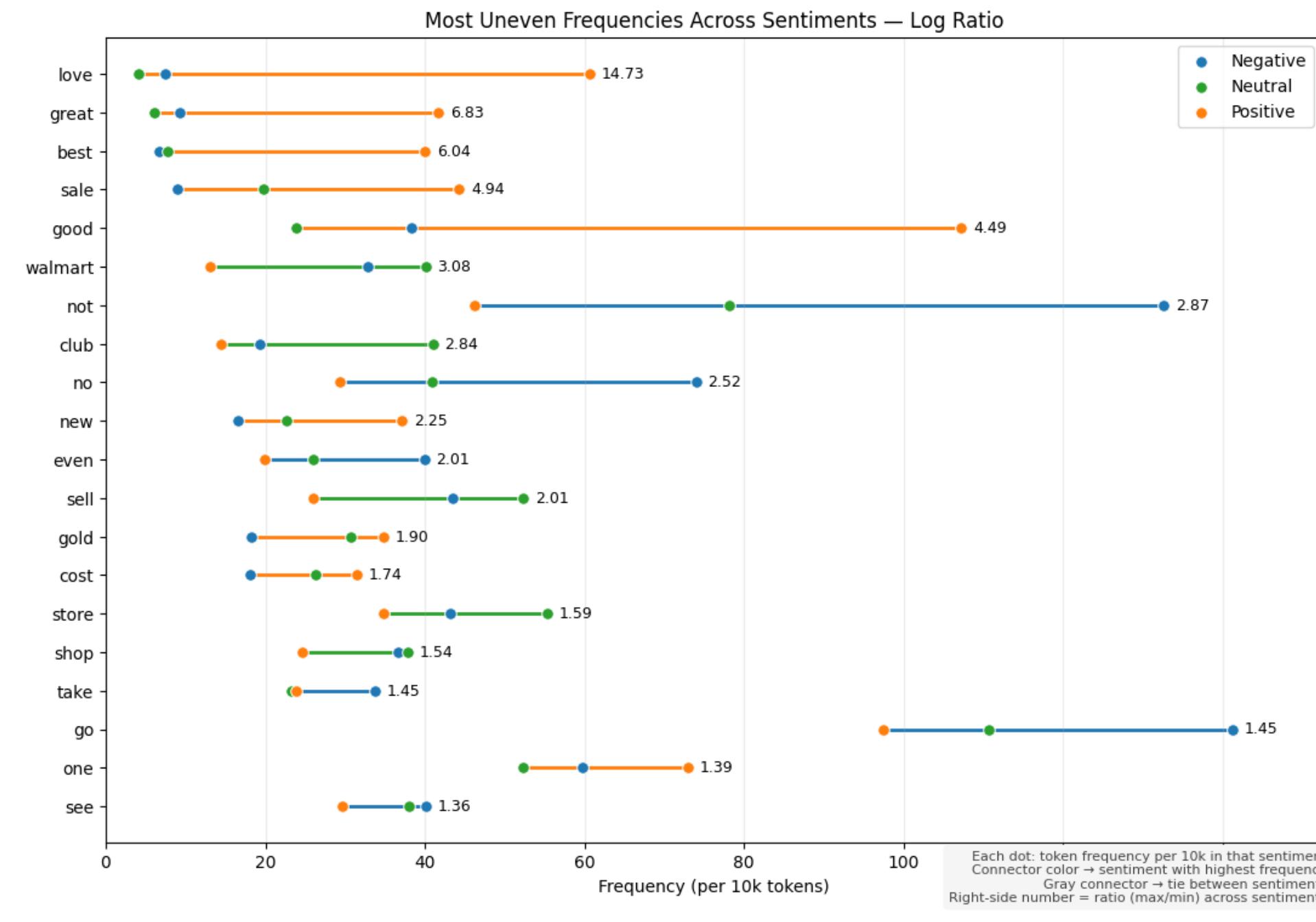
- Sentiment: roBERTa pretrained transformer
- Topics: Latent Dirichlet Allocation



## Results - Domain Sampling: Costco



Content	@Ronxyz00 Case of coconut water was 10 last week at Costco. Today was 12.79
Content_clean	case of coconut water was 10 last week at costco today was 12.79
text_for_sent	@USER Case of coconut water was 10 last week at Costco. Today was 12.79
tokens	case of coconut water was last week at costco today was
lemmas	case of coconut water be last week at costco today be
no_stop	case coconut water last week costco



## LLM Extracted Themes

### Costco hot dog as value symbol

- Hot dog price/changes as shorthand for “is Costco still on our side?” and overall value perception.

### Store choice & family/bulk tradeoffs

- Cross-shopping Costco, Trader Joe’s, BJs, etc.
- Parents weighing bulk value, convenience, and “healthy enough” options for kids.

### Price, protein, and quality concerns (esp. chicken/meat)

- Repeated focus on price vs. quality/safety of staples (rotisserie chicken, meat, etc.).
- Reliability and transparency here act as trust levers.

### Costco vs. Sam's Club performance

- Direct comparisons on stock, electronics, and service.
- Perception is shaped by concrete outcomes (what’s on shelves, issue resolution).

### Membership ROI, everyday essentials

- Joining/renewing justified by savings on batteries, tools, pantry staples, etc.
- Small frictions in staple items can tip sentiment on overall membership value.

### Niche products, ethics, and lifestyle

- Interest in halal/healthier options, protein sources, specialty goods.
- Signals whether Costco aligns with specific dietary, cultural, or lifestyle needs.

## Conclusions

- Co-domain approach dilutes topic interpretability but contributes additional key word context for niche topics.
- Raw frequency yields additional insights when paired with ranked IDF metrics.

## References

- Herk, N. (2025, March 14). *How to Actually Scrape Twitter/X Data with n8n* [Video]. YouTube. <https://www.youtube.com/watch?v=lEo7I-Agj0UY>
- Starmer, J. (2020, March 18). *Latent Dirichlet Allocation* [Video]. YouTube. <https://youtu.be/T05t-SqKArY?si=ZxP2DZt9vDicqIsc>

## Acknowledgements

- DOED MSEIP Grant#: P120A220015, “Boosting STEM Student Success”, Dr. Bernadette Hence at U.S. DOED.
- Dr. Mary Parker, PI Project Director MSEIP.