Text Analysis: Health Misinformation

Sharon Jepkosgei
Christopher Womble
Kendall Goodland

Background and Literature

- 80% of people online use the internet to search for health information yet it's plagued with fake news.
- Many studies on fake news in other domains but few on health sector.
- Those in health domain are mainly COVID-19 specific.
- Our study – Understand health misinformation in a more general sense.

Research Questions

1. Does health misinformation exhibits distinct language patterns, with certain words appearing more than other?

2. Does health misinformation have high prevalence of certain topics?

Research Hypothesis

1. Health misinformation exhibits distinct linguistic patterns, with certain words recurring more frequently than others.

2. Health misinformation has distinct thematic patterns, with certain topics appearing more than others.

Data

Source

- HealthNewsReview.org.
- Health publication reviewed by experts

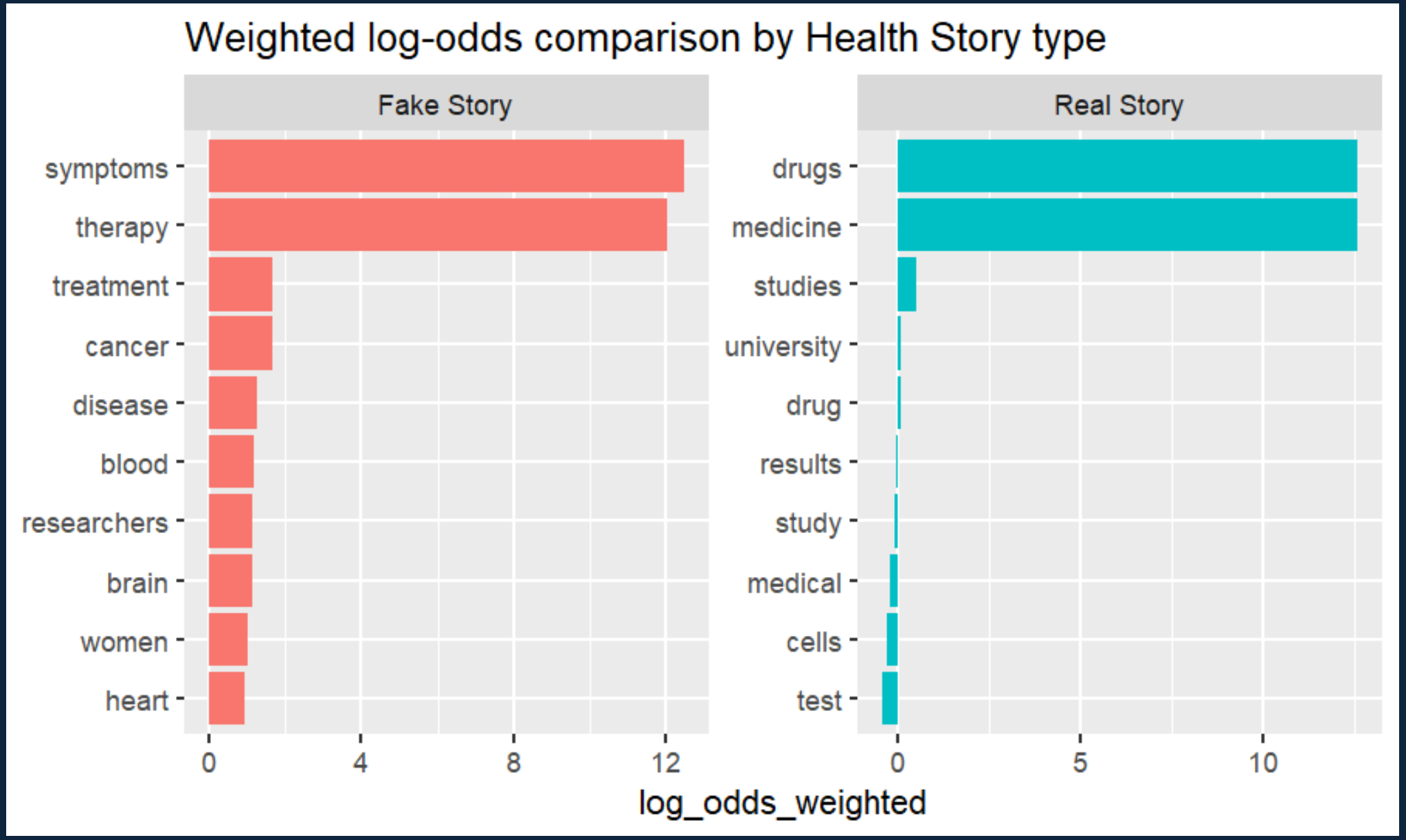
Sample Size

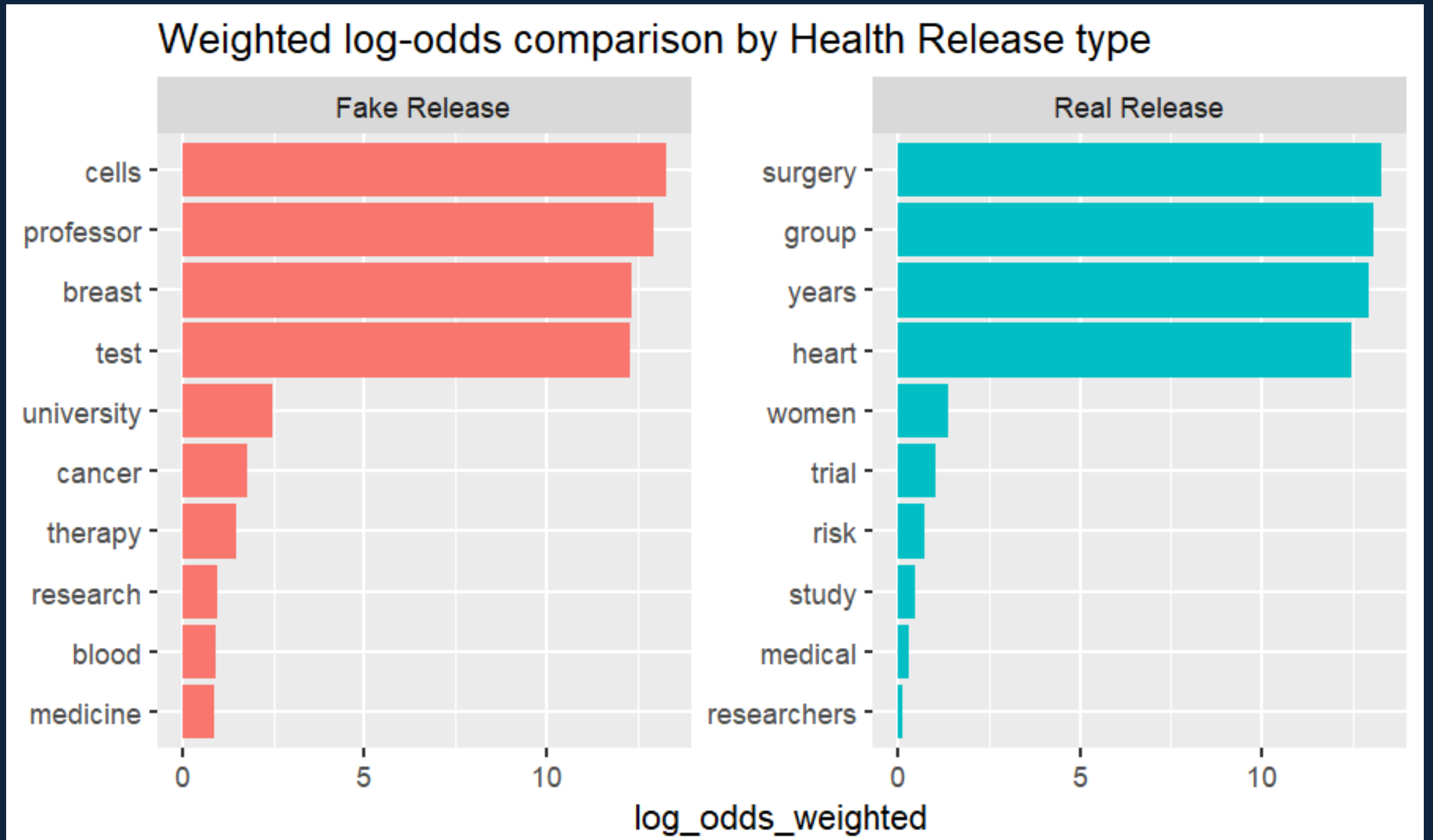
- 1591 documents
- 596 –health release
- 995- health stories

Credibility Criterion

- Less than 3 = fake health information

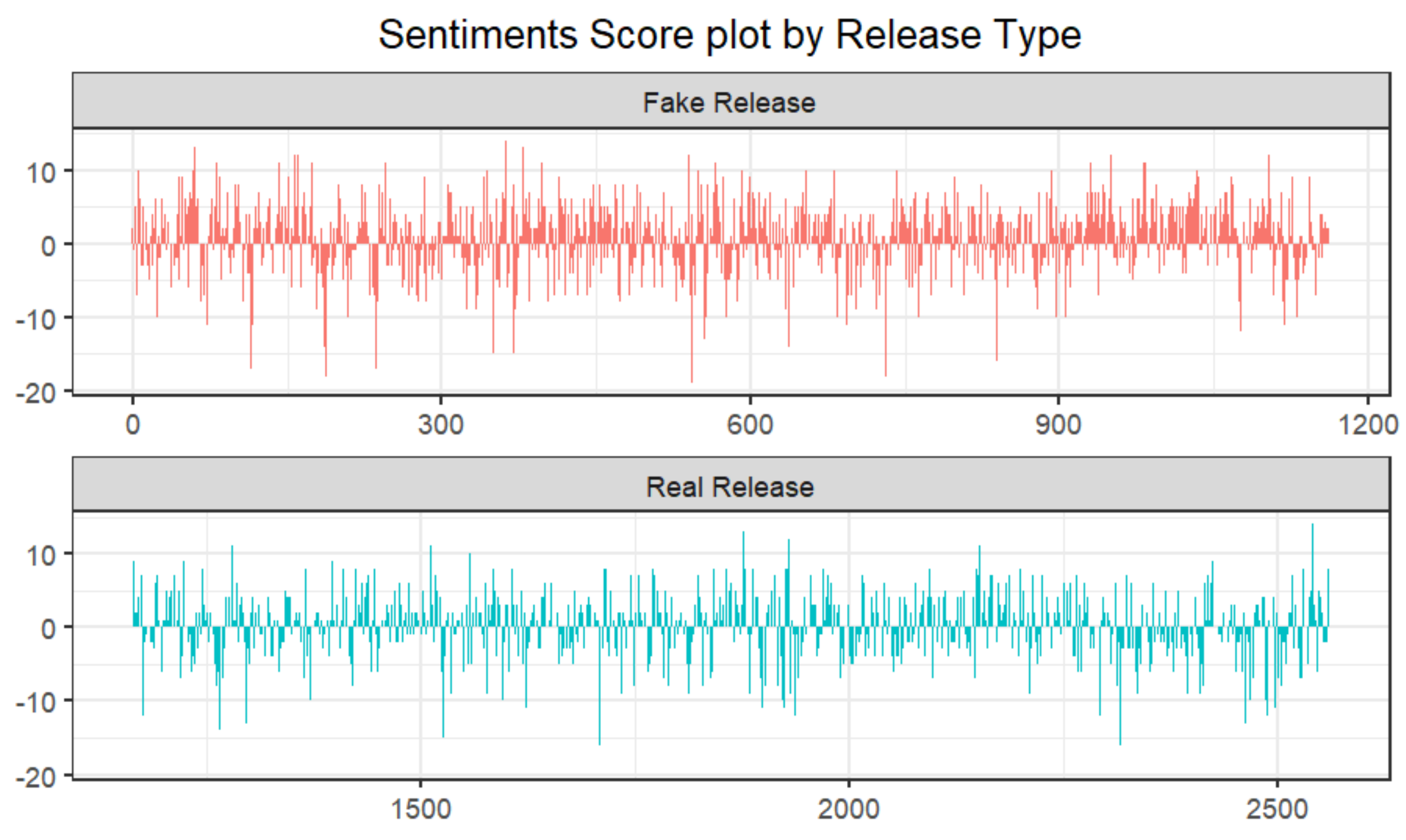
Methods and Results: Word Ratio Analysis

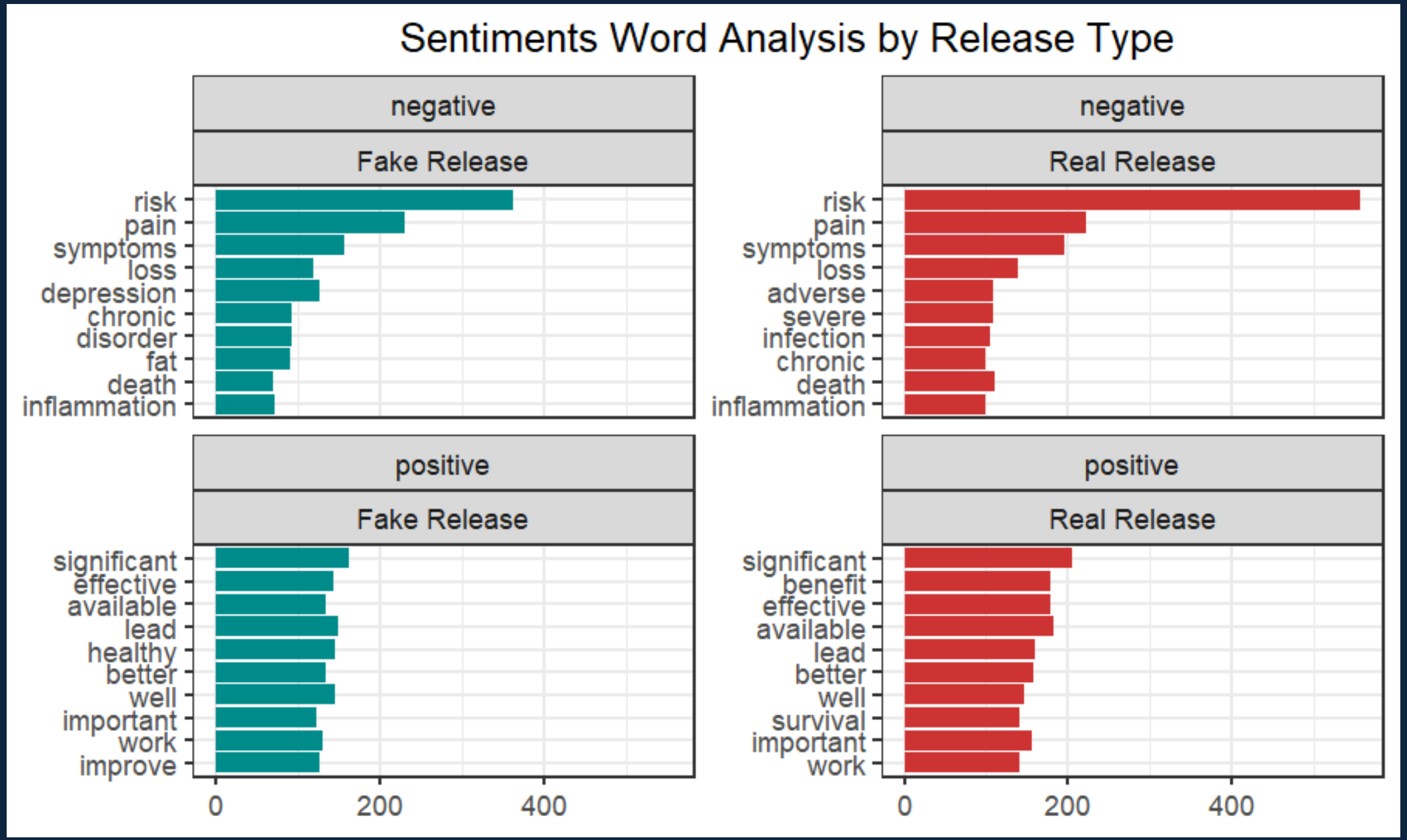



Weighted log-odds comparison by Health Release type
Fake Release
Real Release
cells
professor
breast
test
university
cancer
therapy
research
blood
medicine
surgery
group
years
heart
women
trial
risk
study
medical
researchers
0
5
10
0
5
10
log_odds_weighted

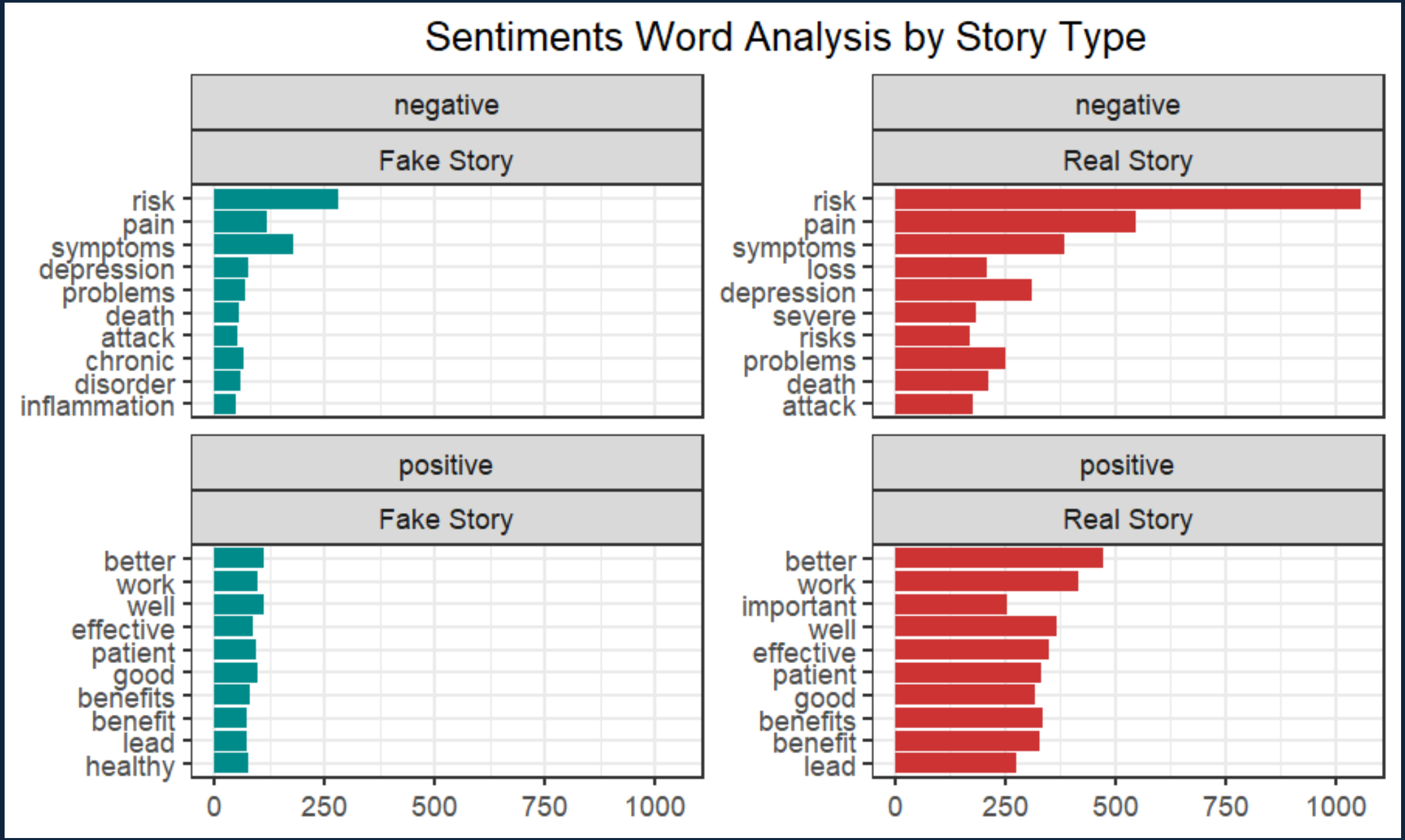
Sentiment Analysis/ Opinion Mining

- Sentiment analysis, a form of NLP that evaluates data for neutrality, positivity, or negativity.
- It involves considering individual words and aggregating their sentiments.
- The tidytext package provides access to various sentiment lexicons
- Bing lexicon for emotional intent analysis.
- Bing, structured around single words, assigns negative or positive sentiment to many English words.


Sentiments Score plot by Release Type
Fake Release
10
0
-10
-20
0
300
600
900
1200
Real Release
10
0
-10
-20
1500
2000
2500


Sentiments Word Analysis by Release Type
negative
Fake Release
risk
pain
symptoms
loss
depression
chronic
disorder
fat
death
inflammation
negative
Real Release
risk
pain
symptoms
loss
adverse
severe
infection
chronic
death
inflammation
positive
Fake Release
significant
effective
available
lead
healthy
better
well
important
work
improve
positive
Real Release
significant
benefit
effective
available
lead
better
well
survival
important
work
0
200
400
0
200
400


Sentiments Score plot by Story Type
Fake Story
10
0
-10
-20
0
200
400
600
Real Story
10
0
-10
-20
1000
2000
3000


Sentiments Word Analysis by Story Type
negative
Fake Story
risk
pain
symptoms
depression
problems
death
attack
chronic
disorder
inflammation
negative
Real Story
risk
pain
symptoms
loss
depression
severe
risks
problems
death
attack
positive
Fake Story
better
work
well
effective
patient
good
benefits
benefit
lead
healthy
positive
Real Story
better
work
important
well
effective
patient
good
benefits
benefit
lead
0
250
500
750
1000
0
250
500
750
1000

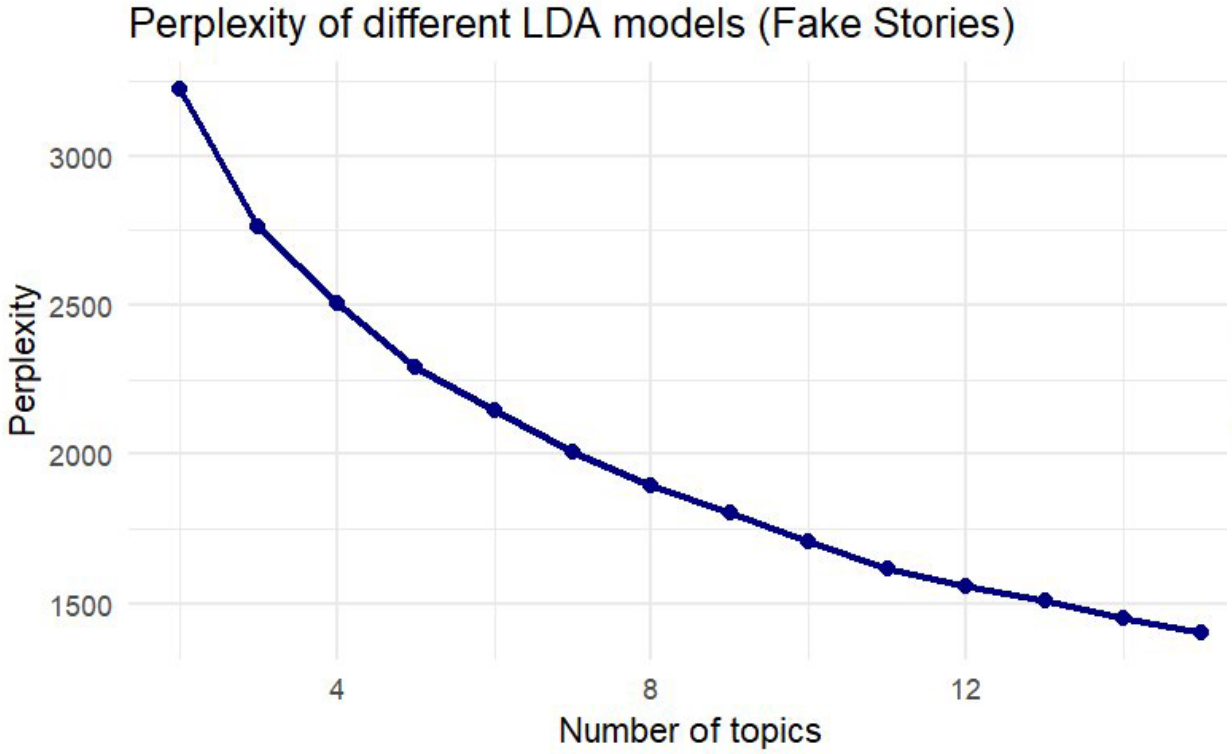
Topic Modeling

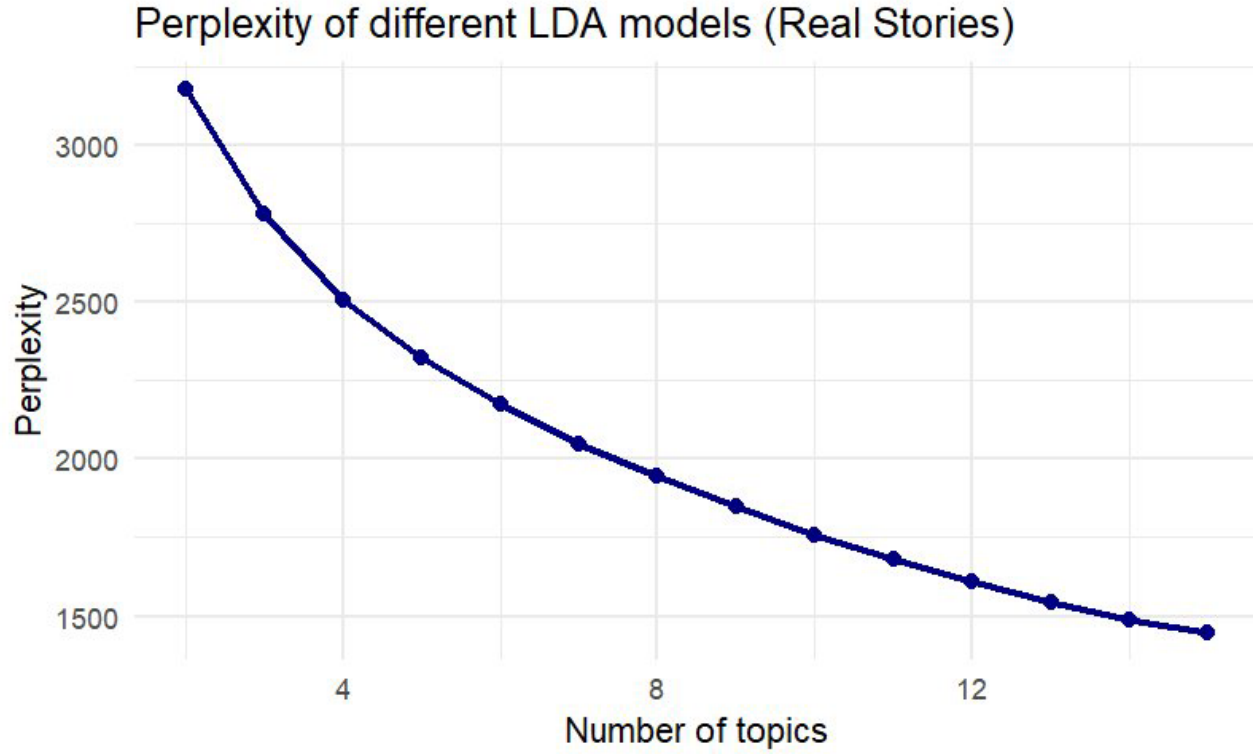
Topic modeling for finding and tracing clusters of words in large bodies of texts.

Latent Dirichlet Allocation for its ability to better generalize and its popularity in the humanities fields

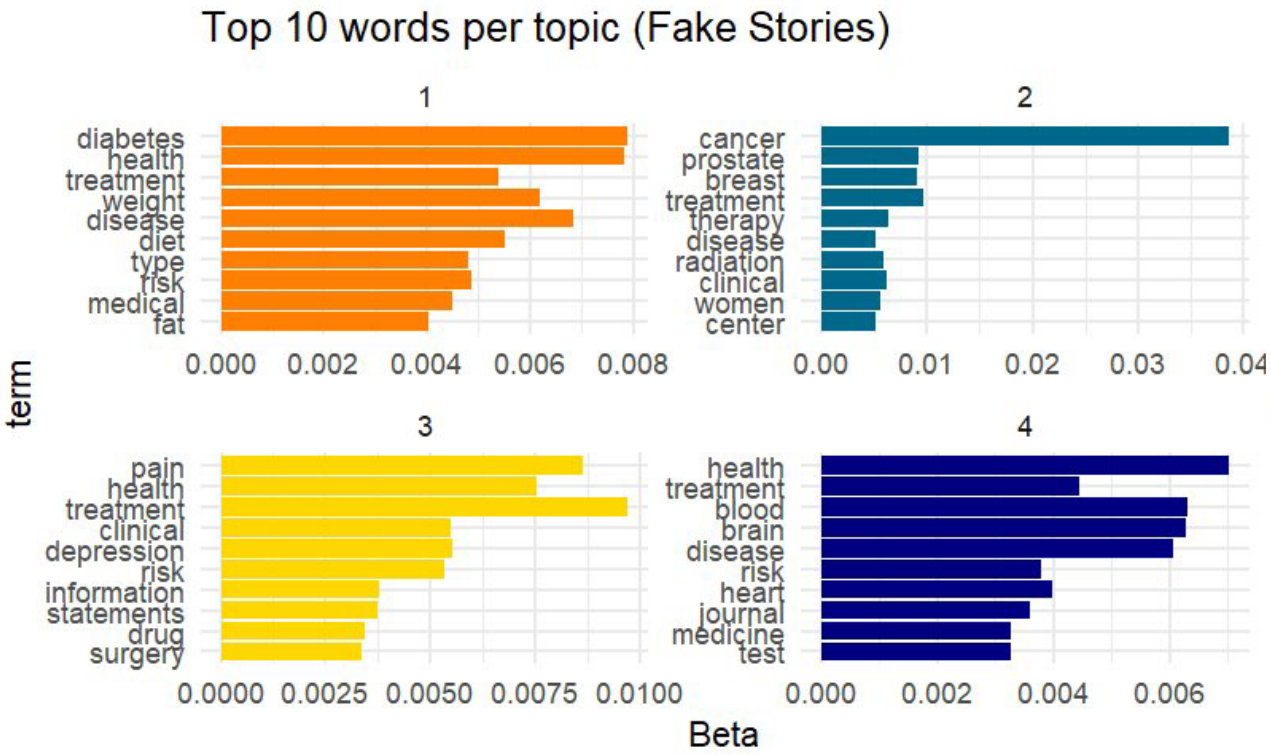
Used elbow plots that mapped the coherence (AKA the perplexity) against the number of topics.

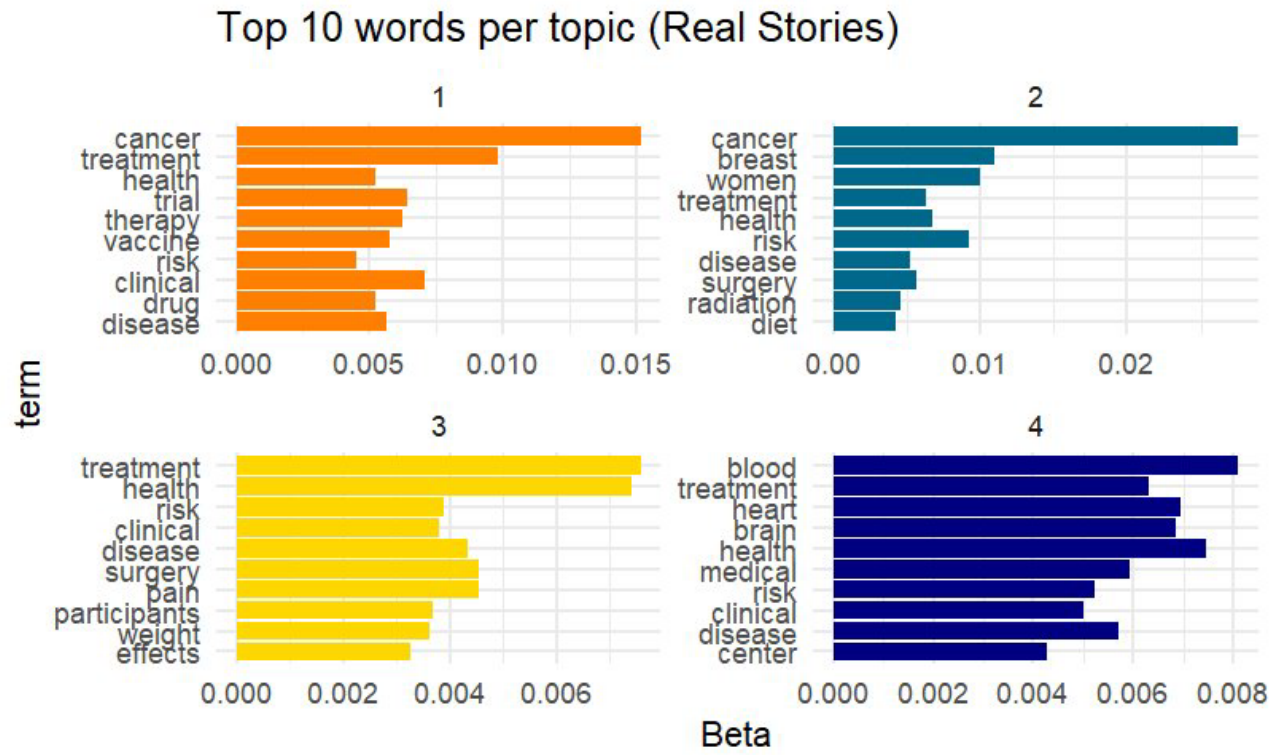
Health Release



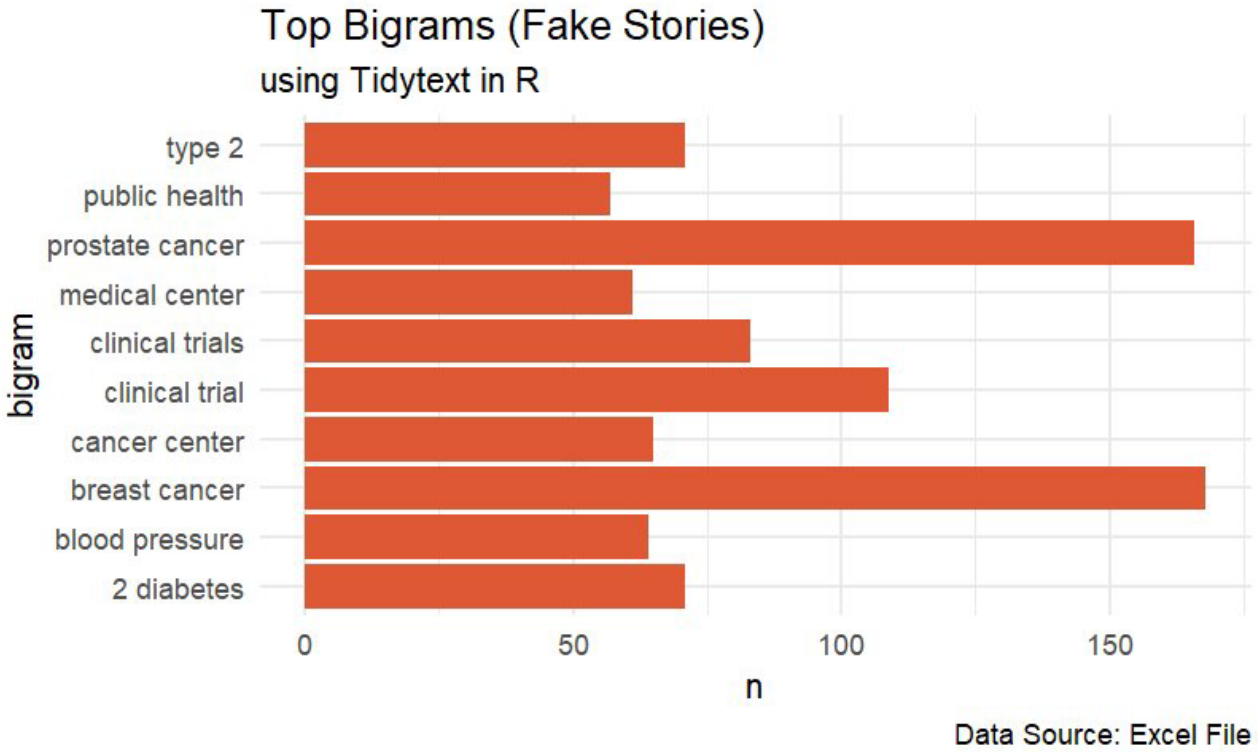


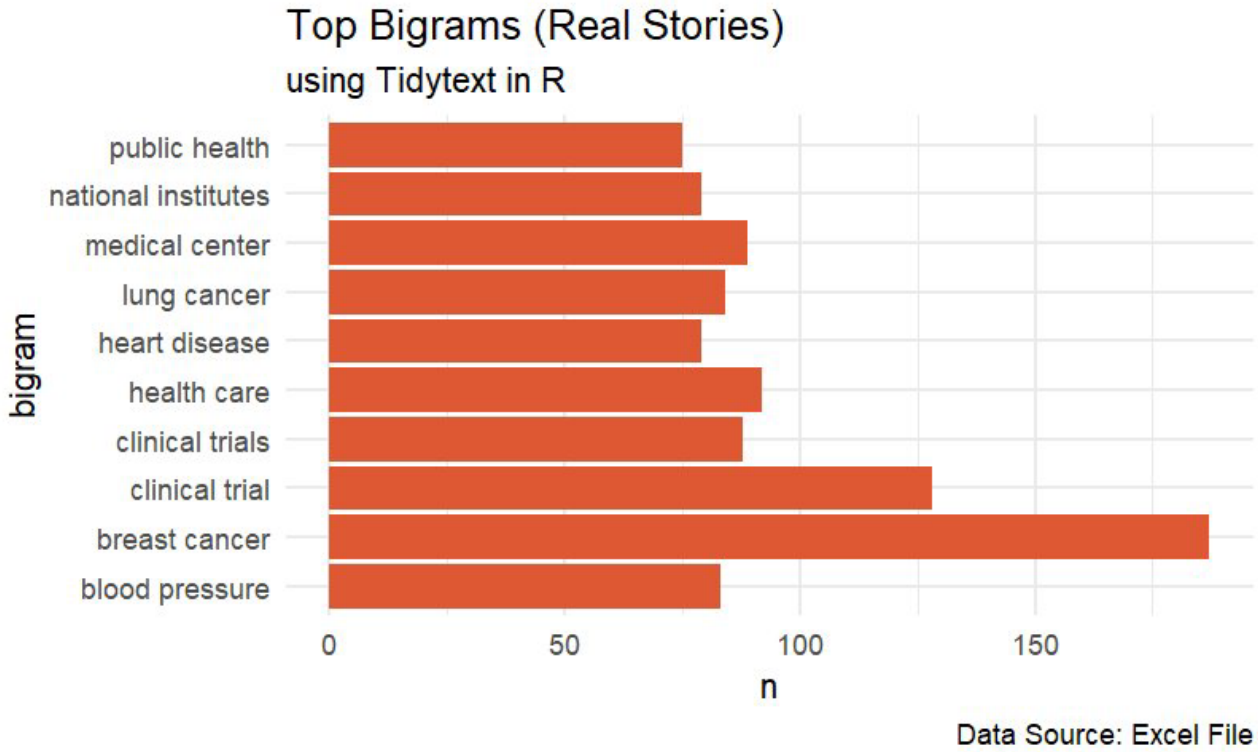
Health Release



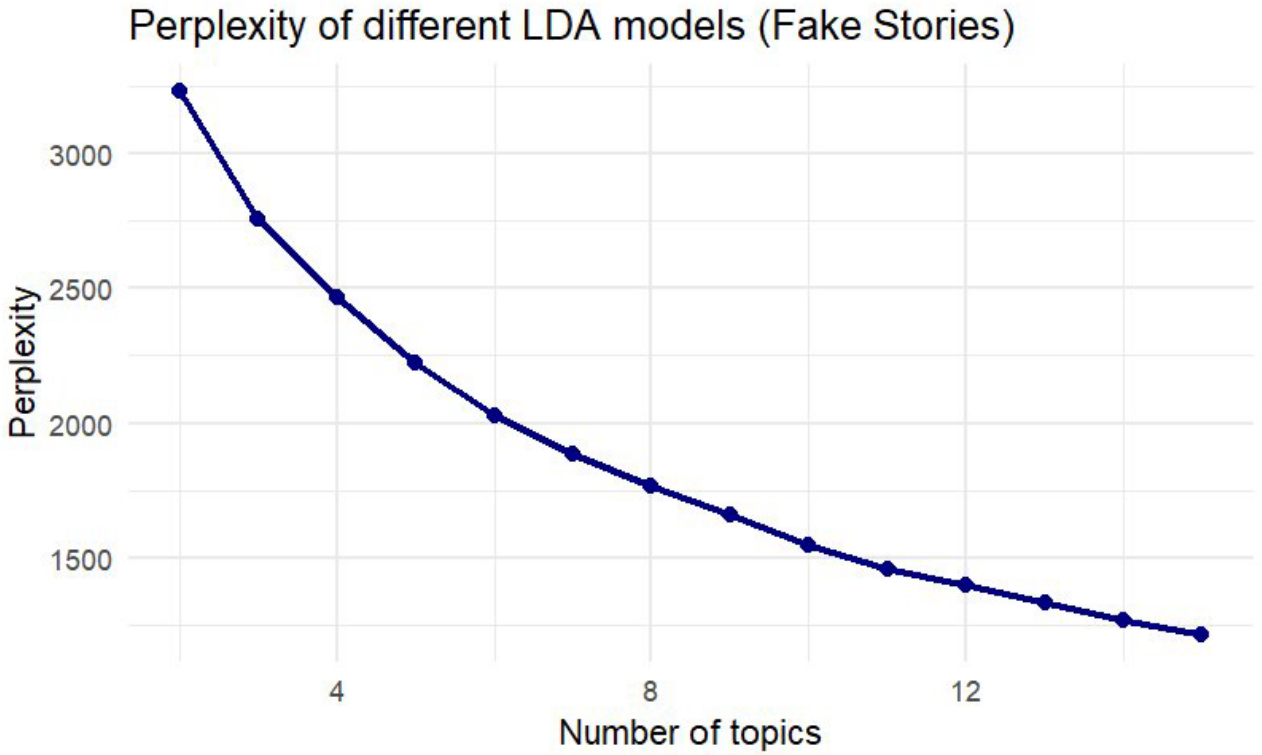


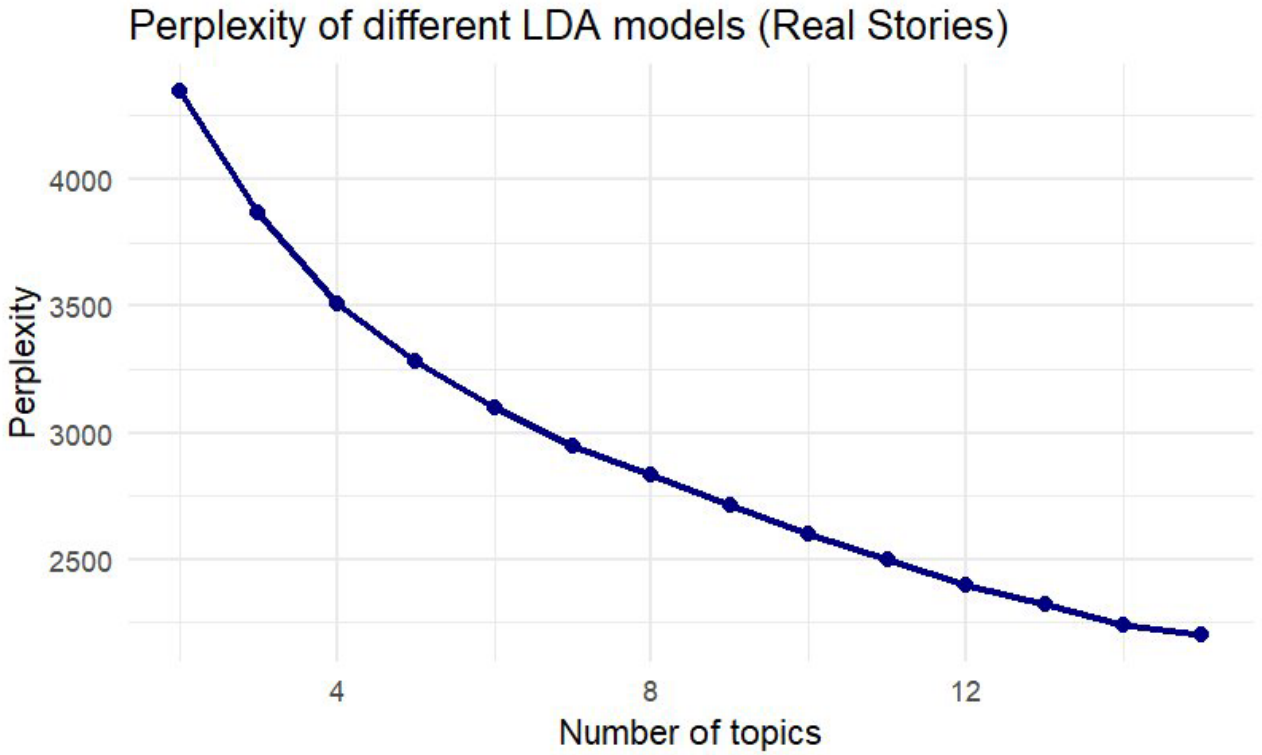
Health Release



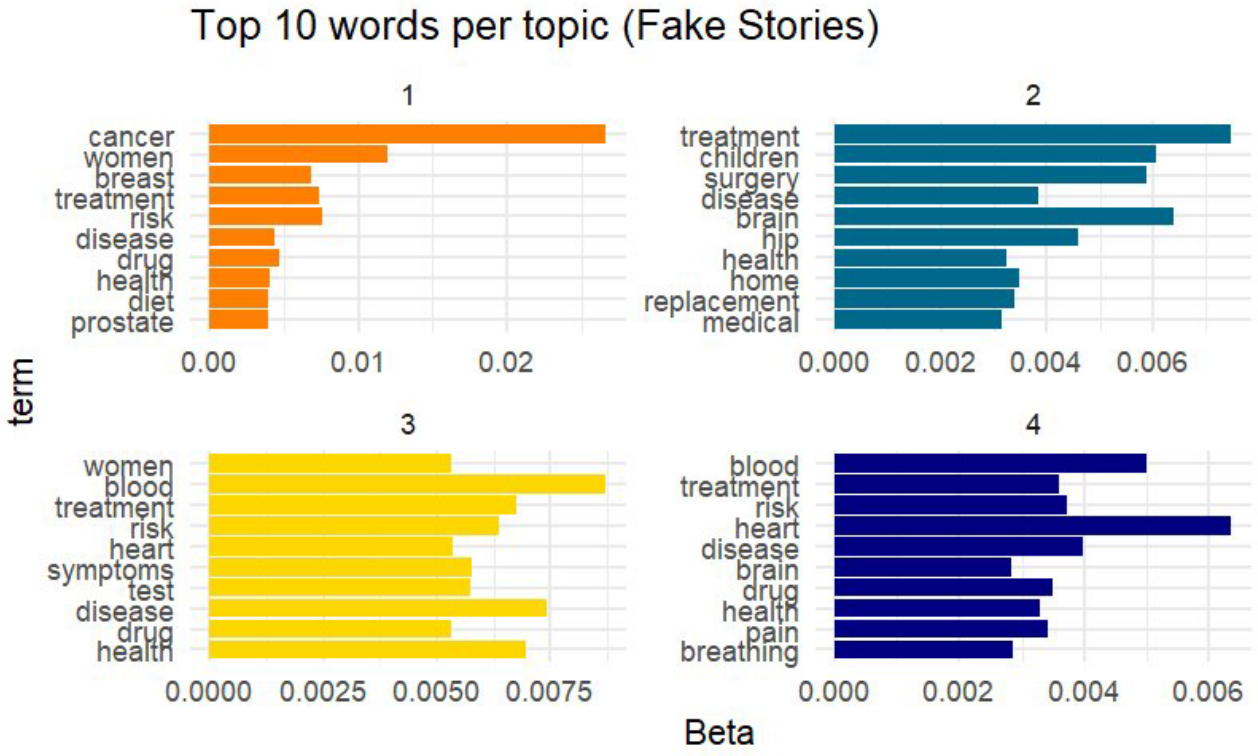


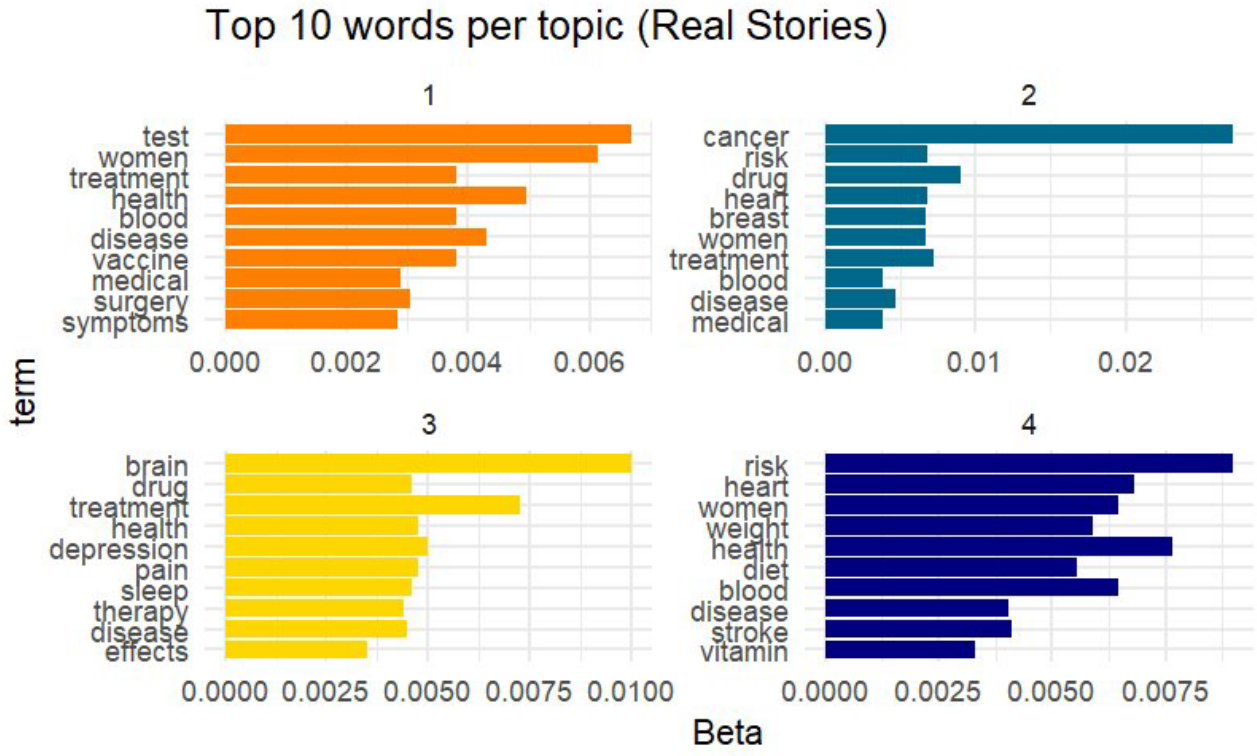
Health Stories



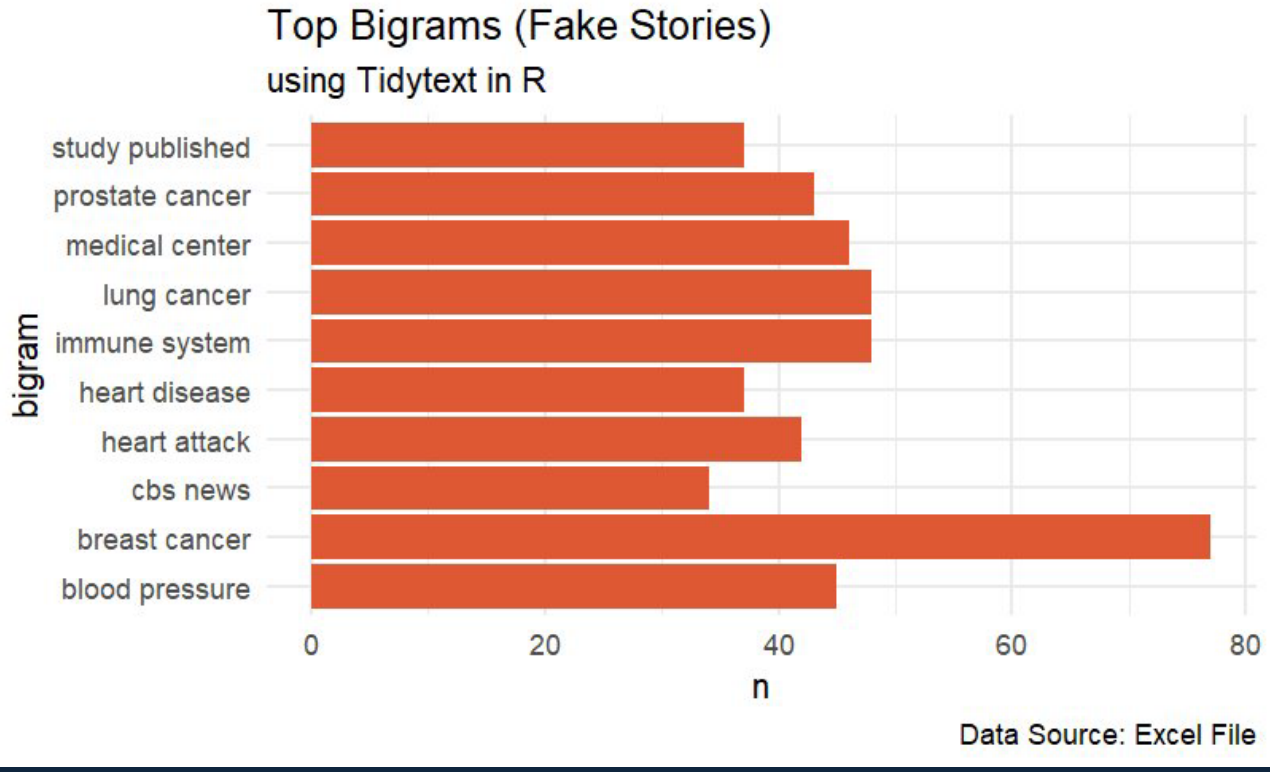


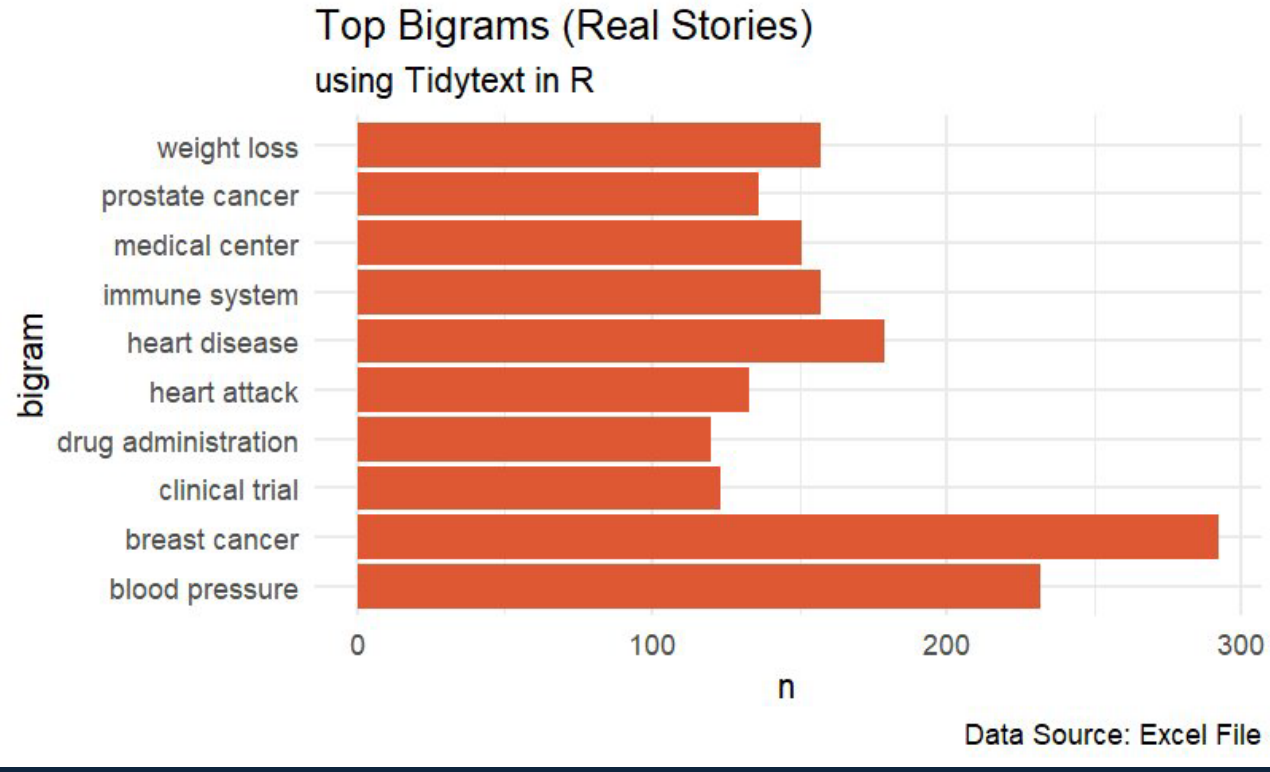
Health Stories





Health Stories





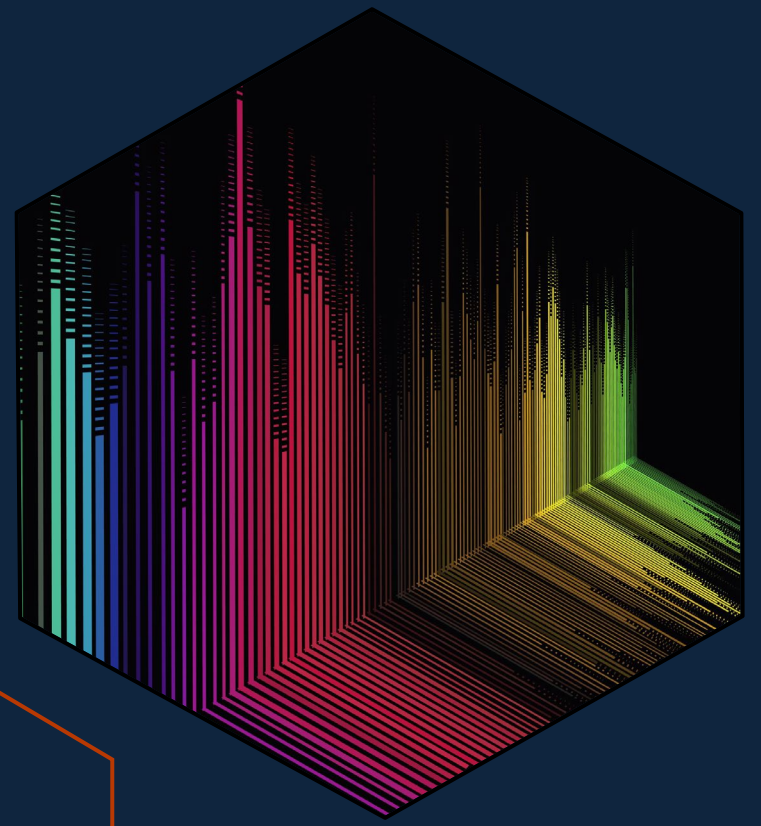
Takeaways and Future Research

1. Mixed support for both our hypotheses

- Some differences in language use between fake and credible health publications.
- Some distinct topics in fake versus real health articles

2. Ideas for future research

- Use a larger sample of documents from variety of sources and conduct a deeper qualitative research
- Compare other sources of health information such as social media






Thank you