Sharon Jepkosgei, Christopher Womble, Kendall Goodland

EPPS 6323: Knowledge Mining

Dr. Karl Ho

May 3rd, 2024

Health Misinformation/Disinformation: Common Themes, Words, and Patterns

Introduction

In recent years, the proliferation of health misinformation has become a significant concern, with the potential to undermine public health efforts and contribute to misconceptions about medical treatments, diseases, and preventive measures (Wang et al., 2019). To address this issue, it is essential to conduct a comprehensive text analysis to identify common themes, words, and patterns prevalent in fake health information (Zhou et al., 2020, Hayawi et al., 2022). This project's main objective is to analyze a corpus of health-related texts to identify recurring themes, prominent words, and dissemination patterns associated with misinformation. By doing so, we aim to gain insights into the prevalent forms and sources of health misinformation, enabling targeted interventions and educational campaigns to counteract its spread. This focuses on two main research questions: (1) Does health misinformation exhibit distinct linguistic patterns with more words appearing more than others? and (2) Does health misinformation have distinct thematic patterns, with certain topics appearing more than others?

To address these questions, we collect a corpus of 1591 publications of health stories and health-related publications from universities and other research institutions. We use numerous text analysis techniques including topic modeling, sentiment analysis, word ratio analysis, bigram analysis, and word frequency analysis. We find some evidence that health misinformation does have some distinct linguistic and thematic patterns.

Review of Literature

Health misinformation is not a new issue and has demonstrated how essential accurate health data and news is in daily life. The legal aspect of this phenomenon is also cloudy. Most of the time, it is easy to classify misinformation as “speech” – this leads to a lack of legal repercussions for bad actors who maliciously spread misinformation online and through other mediums. These issues became glaring during the onset of the COVID-19 pandemic. Many odd theories and remedies regarding the virus were in the spotlight on social media, including claims that China purposely created COVID-19 as a bioweapon, and that 5G radiation was causing the virus (Goodyear, 2022). In acknowledging the threat of fake news to public health, researchers have begun the process of finding solutions to the issue. One study focuses on the collection of news data alongside the COVID-19 pandemic and matching them up to quality criteria (Montesi, 2021). This is a popular approach in which researchers gather data and sort it into categories based on quality indicators, such as falsifiability and reputability (Montesi, 2021).

The strategy of topic modeling has been utilized in many ways, specifically with social media and text data. One study gathered Twitter data with information regarding the COVID-19 pandemic and cleaned the data based on relevant keywords (Lyu et al., 2021). The data gathered was tracked across the COVID-19 timeline and results indicated a positive association with social media and public trust (Lyu et al., 2021). Another study using topic modeling and sentiment analysis also revealed that those strategies were effective in producing a positive relationship between vaccine rollout and vaccine trust (Huangfu et al., 2022). What is interesting is that both studies collected data on COVID-19 sentiments around the time of the first wave of vaccine rollouts. This aptly improved the data processing because researchers were able to use key words like “vaccine” and “vaccinations”. Content is clearly relevant to sentiment when it comes to news and media, especially public health. These two studies focused on the pandemic period and are limited in its generalization. Our current research is focused on patterns in health misinformation before and after the COVID-19 pandemic.

Research Hypotheses

First Hypothesis

H01: Health misinformation has no distinct linguistic patterns, with certain words recurring more frequently than others.

HA1: Health misinformation exhibits distinct linguistic patterns, with certain words recurring more frequently than others.

Second Hypothesis

H02: Health misinformation has no distinct thematic patterns, with certain topics appearing more than others.

HA2: Health misinformation has distinct thematic patterns, with certain topics appearing more than others.

Methodology

Data

The source of our data is HealthNewsReview.org. It is free from industry and supported by Laura and John Arnold Foundation and Informed Medical Decisions Foundation. HealthNewsReview.org reviews health publications from news releases by organizations or research institutions and main US media outlets. For the final project, we collected a corpus of 1591 documents, 995 health stories, and 96 health releases. The stories are publications on efficacy claims about specific treatments, tests, products, or procedures discussed in media outlets such as CNN. The health releases are publications by companies, universities, or other research institutions. These articles we collected have been assessed based on a standard rating system, in which each document is reviewed by at least two reviewers with years of experience in the health care field. The reviewers are all well-versed in fields such as journalism, medicine, health service research, public health and patient health, and each of them signs an industry-independent disclosure agreement. The diversity and independence of the reviewers are expected to mitigate the effects of any political bias in the assessments.

We cleaned our datasets and separated them into fake and real based rating scores. For data from HealthNewsReview.org the rating score ranges from 0 to 5. We adopted the strategy in (Shu et al. 2018), we treat articles whose scores lower than 3 as fake news.

Methods

Code for our techniques can be found here:

https://github.com/sharonjepkosgei/knowledgemining/tree/main/project/code

Word Frequency Analysis and Word Ratio Analysis

Word Frequency Analysis is a commonly used technique in NLP where text is converted into a set of words. In this case it serves two main purposes: to allow comparison across types of publication and identify occurrences of specific words among fake and real health information.

Word ratio analysis builds on the word frequency analysis approach described in the previous section. It explores the comparison of the frequency of specific words in fake and real publications through editorial leanings. By calculating word ratios, this analysis aims to show potential biases and thematic differences in how the health information is presented in both fake and real publications. The basis of the analysis is the calculation of word ratios. First, through term frequency we compared the frequency of some words in fake and real health stories as well as health releases. This method was applied separately for each type of publication, release, and story. The comparative analysis focused on identifying words that were more prevalent in credible versus fake health news in both stories and releases. It is important to highlight differences in language use in fake and real health information. Another word ratio analysis we conduct is the term frequency-inverse document frequency and log odds ratios which assess the significance of a word within a specific document compared to the entire corpus by accounting for frequency of the word within the category and its rarity across other groups. The results of the word ratio analysis are presented in a series of tables, showing the top words with the highest ratios in each category of health information and type of publication. These tables are a good visual representation of the difference in language use between fake and real health information, but they only offer qualitative insights and lack quantitative depth.

Sentiment Analysis

The next technique we apply in our analysis is sentiment analysis, or opinion mining. It is an NLP tool often used in text analysis to assess data neutrality, positivity, and negativity. In our analysis we use sentiment analysis to investigate the presentation of health information in credible and fake publications. A common way to do sentiment analysis is to evaluate text as a combination of individual words and obtain the overall sentiments. The tidytext package avails several free sentiment lexicons used for analysis the emotions and opinions in text data. For our analysis, we used the *bing* lexicon. It is a technique that provides the emotional intent of words in a series of text. *Bing* is a dictionary based on unigrams and contains English words that have been assigned a negative or positive sentiment.

For sentiment analysis, the text was extracted from both the health releases and health stories. The text was then converted into tidy text format (one-token-per-row) using unnest_tokens(). The bing lexicon serves the role of categorizing words from the publications into binary clusters, signifying negative or positive sentiments. The R Scripts provided show the procedure of sentiment analysis. Nevertheless, the key functions used are get_sentiments() which avails the sentiment lexicons and their measures, inner_join() which calculates sentiment in various ways(), and mutate() which determines net sentiment in each text section. Through sentiment analysis

Topic Modeling

Topic modeling is a well-known and sound method of making sense of vast amounts of text data, and it does this by finding and tracing clusters of words in order to see what sort of underlying story is being told in the data. In our project, we built Latent Dirichlet Allocation (LDA) models in R, largely due to their ability to better generalize data and due to their popularity in disciplines focused on humanities in general. We cleaned and separated the release and story data, then created document-term matrices for both the real and fake data separately, and then created LDA models for each. With those LDA models, we went on to create elbow plots that displayed the perplexity of the LDA models on the X-axis against the number of topics of the Y-axis, each with four topics. The purpose of this was to see if our model was effectively capturing what we wanted it to, which was the overall coherence of our dataset. Once that was done, we next created models that displayed the top ten words that appeared in all four of the topics assigned to the LDA models, to better grasp the data's content. Seeing the ten most frequent words per topic would allow us to see what sort of patterns appear within release data compared to story data and misinformation compared to real information.

Bi-gram Analysis

Once the top ten models were built, we moved on to the last step, the bigram models. In a similar vein as the top ten models, the bigrams show the top ten pairs of words per topic. The top bigrams can offer better context for the top ten words, allowing us to glean more information about general trends of the language used in our data. Once again, we analyzed the fake and real data separately for both the release and story datasets. With that, we could compare how certain word pairings come about between our real and fake data.

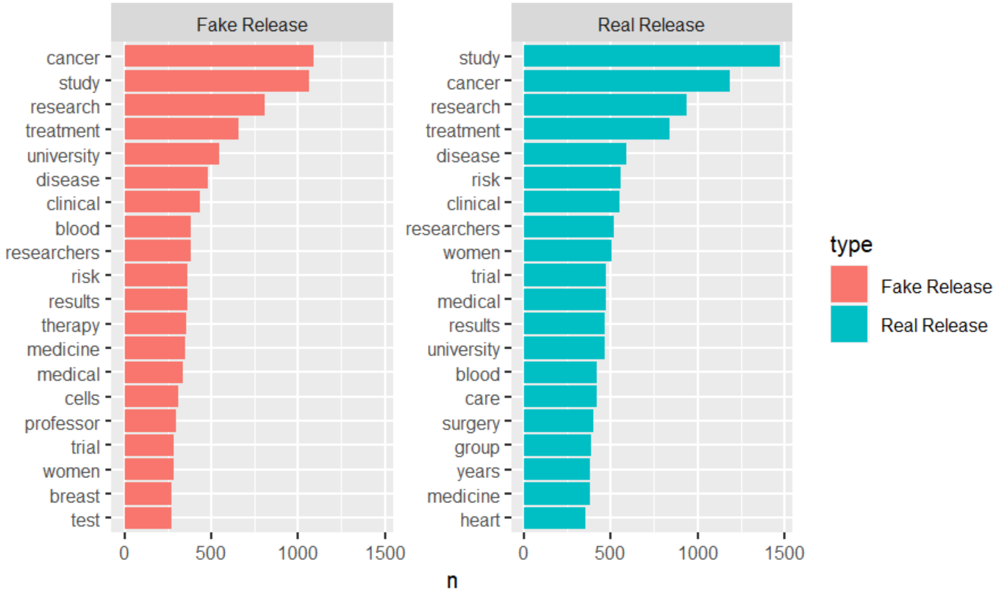
Results/Analysis

Bag of Words/Word Frequency Analysis

Our word frequency analysis reveals a striking similarity in the most frequently occurring words across both fake and credible health news stories and releases. This finding highlights the intricate challenge posed by health misinformation, indicating that distinctions may not primarily stem from the specific words used, but rather from underlying thematic patterns and nuances in language usage.

Figure 1

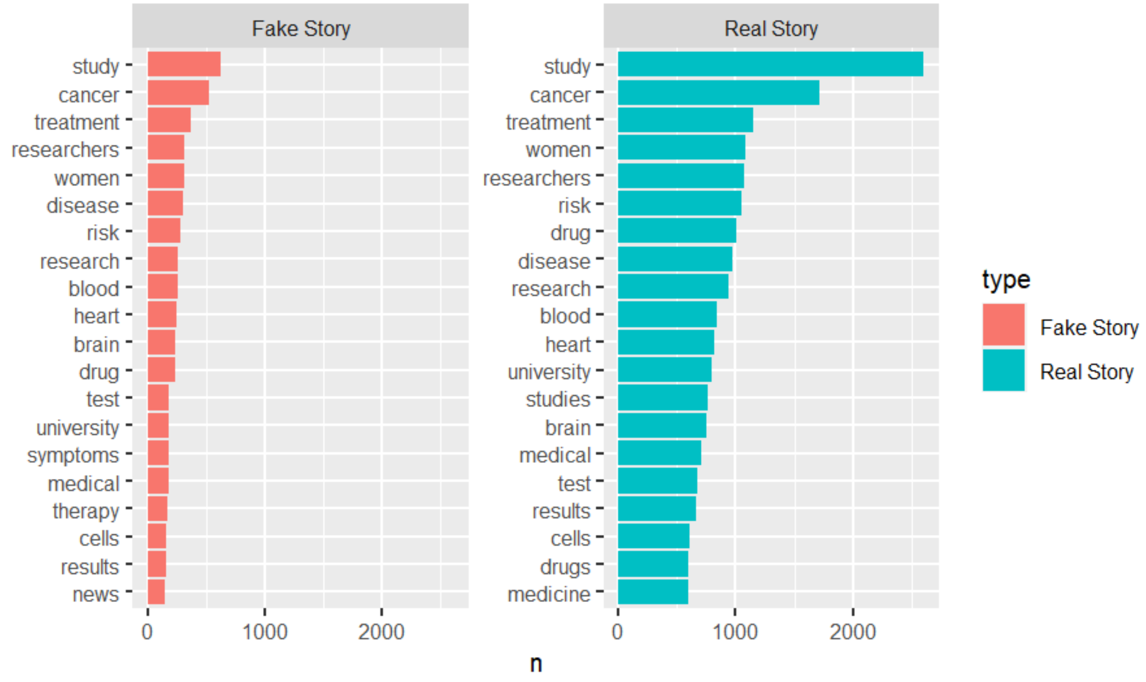
Bag of Words Model for Release Data



This insight prompts a deeper exploration into the subtleties of language employed within health communication. While the surface-level word usage may appear similar between fake and credible sources, it is the contextual framing and narrative structure that likely unveil the true disparities. By delving into these thematic intricacies, we can gain a more nuanced understanding of the mechanisms driving health misinformation and develop more targeted strategies to address its dissemination.

Figure 2

Bag of Words for Story Data



Word Ratio Analysis

Word ratio analysis provides a deeper and more accurate keyword analysis of health misinformation by accounting for sampling disparities in the dataset. The term frequency analysis again provides evidence of challenges of identifying health misinformation because of similar words with the highest ratios in the fake compared to real stories. However, a critical insight from our analysis is a higher proportion of top keywords including “cancer” and “women” in fake health stories compared to credible news.

Figure 3

Term Frequency Model for Story Data

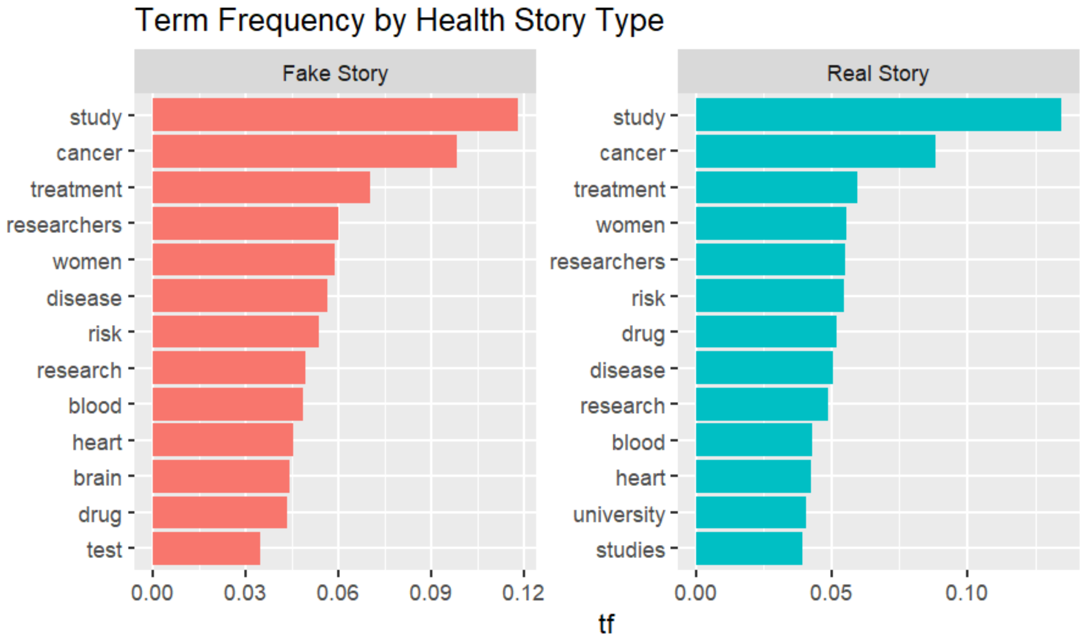
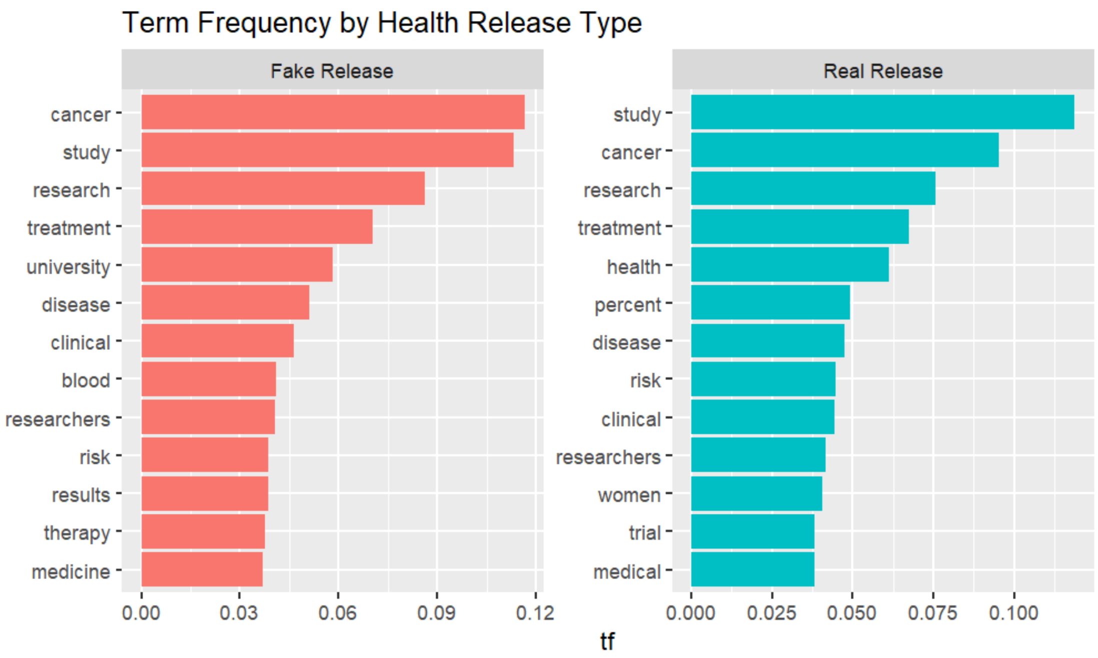


Figure 4

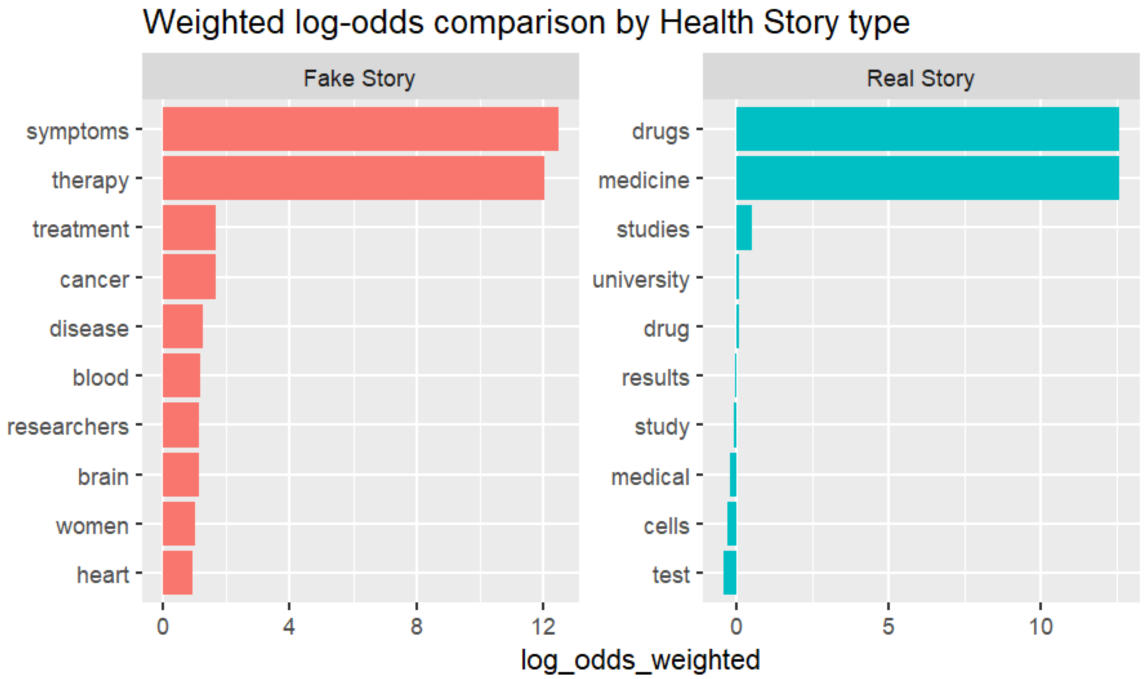
Word Frequency Model for Release Data



We use the log odds ratios to have better insights into the differences in word usage between credible and fake health information. Our analysis shows that the words ‘symptoms’, ‘cancer, ‘women’, and ‘therapy’ are more specific to fake news than credible news.

Figure 5

Weighted log-odds model for Story Data



The log odds ratios analysis of health releases shows that words like ‘symptoms’, ‘cancer, ‘women’, and ‘therapy’ are more specific to fake news than credible news. On the other hand, words like “cells”, “breast”, “test” and “professor” are more likely to appear in fake health releases than highly credible ones. Again, “surgery” and “heart” are words that are more likely to be found within real health releases than those that are fake. It is worth nothing that some words may be more likely to appear in either fake or real health information depending on the type of publication – news or release.

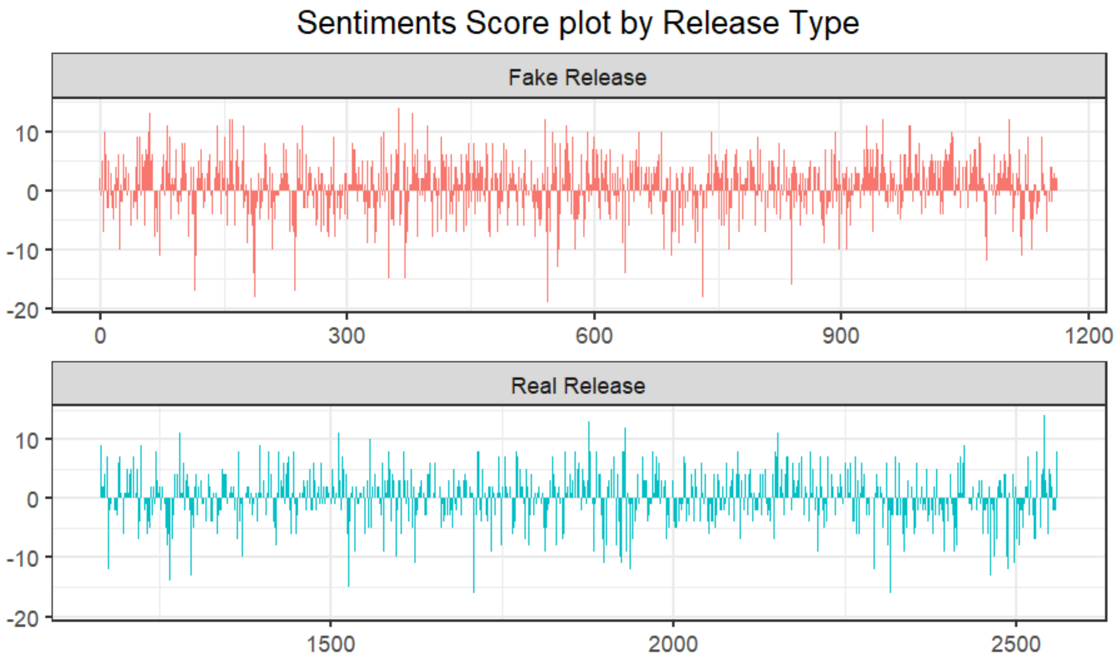
Figure 6

Weighted log-odds Model for Release Data

Sentiment Analysis

Figure 7

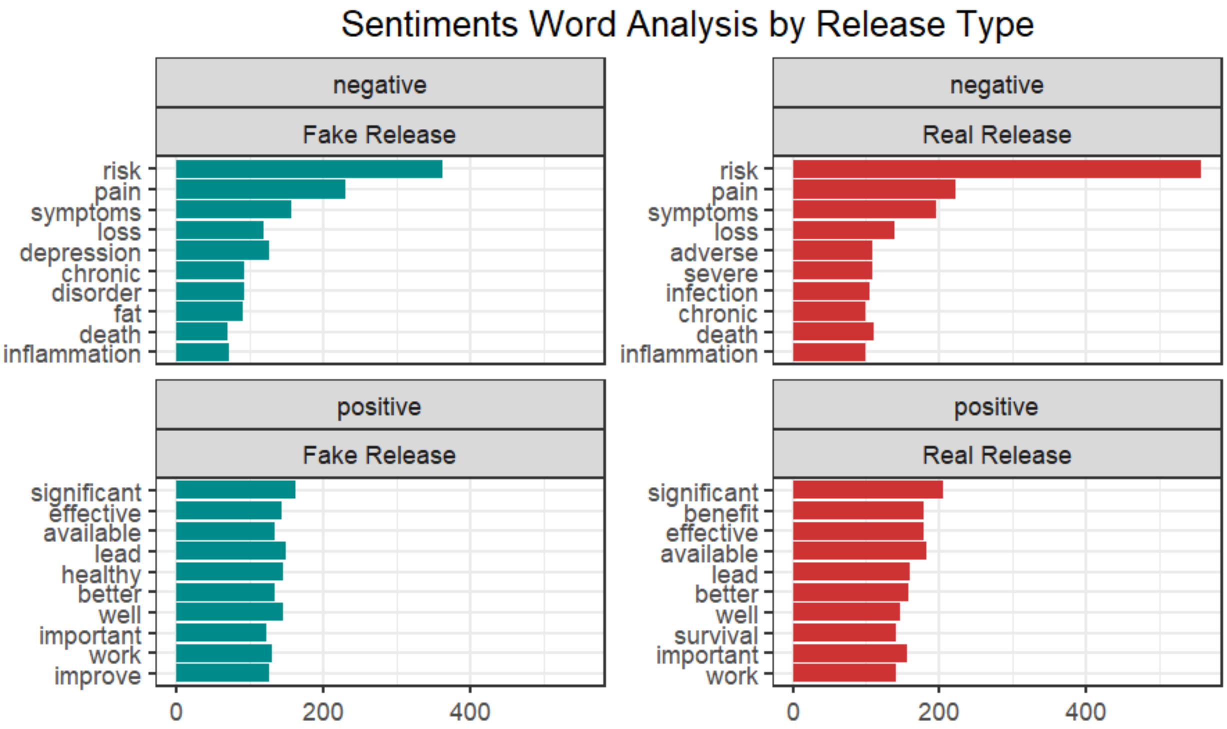
Sentiment Score Model for Release Data



Our *bing* sentiment analysis of health releases provides support for our hypothesis that health misinformation would have distinct language patterns. The analysis revealed notable differences in the sentiment between credible and fake information in health releases. In general, fake releases have a higher proportion and prevalence of negative words compared to credible health publications. This may suggest that articles spreading misinformation in the health sector use language with a negative tone. However, genuine releases show many positive words compared to health misinformation. This disparity implies that credible publications tend to focus on positive health topics.

Figure 8

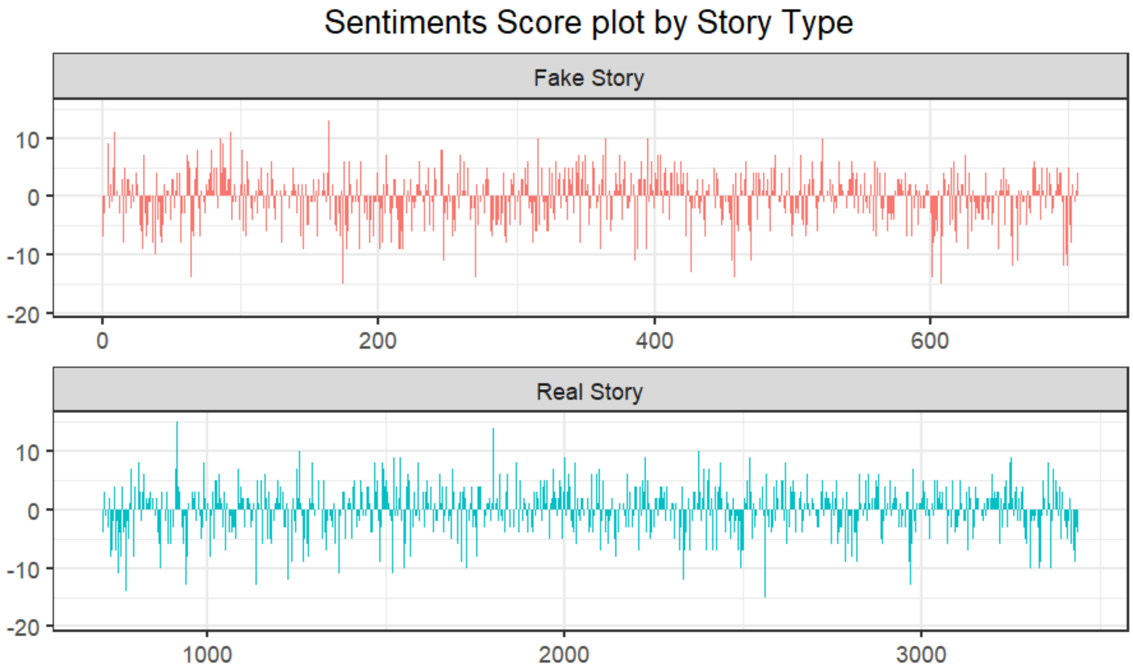
Sentiment Word Analysis Model for Release Data



From the figure, both the positive and negative words highest in prevalence were remarkably similar between the fake and real health releases, suggesting the complexity of identifying language patterns in health misinformation. Thus, we are careful not to overstate the extent to which the sentiment analysis supports our hypothesis.

Figure 9

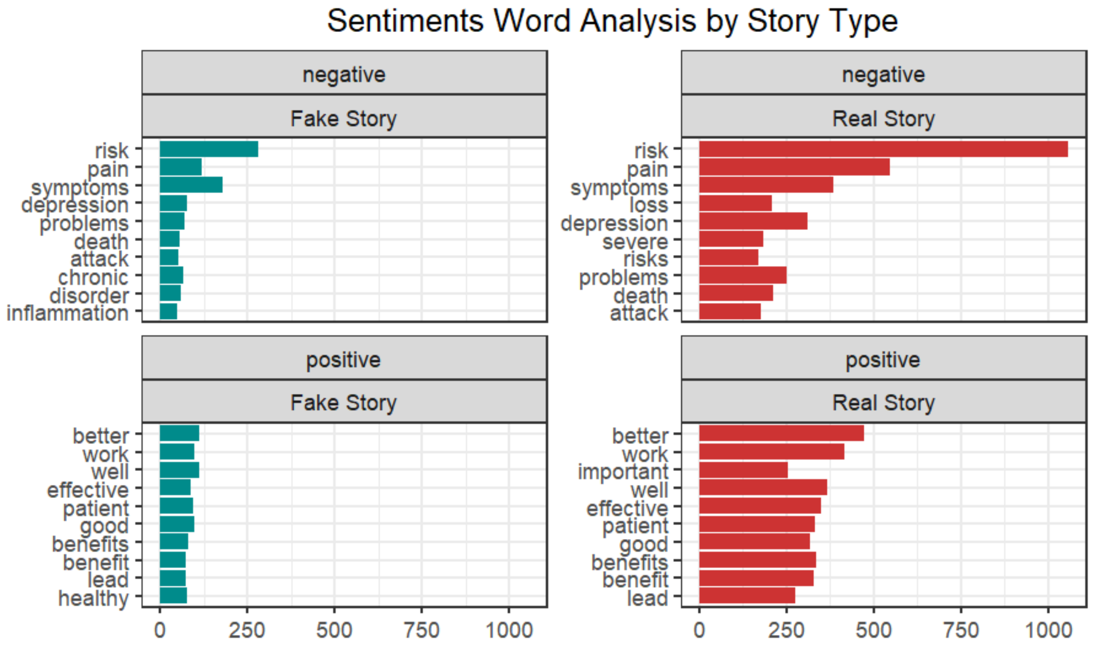
Sentiment Score Model for Story Data



Once more, our bing sentiment analysis unveils a notable discrepancy in tone between fake and credible health information sources. We find that fake health news stories show a high prevalence of negative tone among fake stories compared to credible health news. On the other hand, credible health news has an abundance of positive tone compared to fake news. However, despite these distinct tonal differences, our analysis shows an unexpected convergence in the frequency of top negative and positive words across both types of news stories. Surprisingly, we see similar words appearing in both fake and credible datasets, indicating that the identification of health misinformation may not solely depend on the vocabulary used.

Figure 10

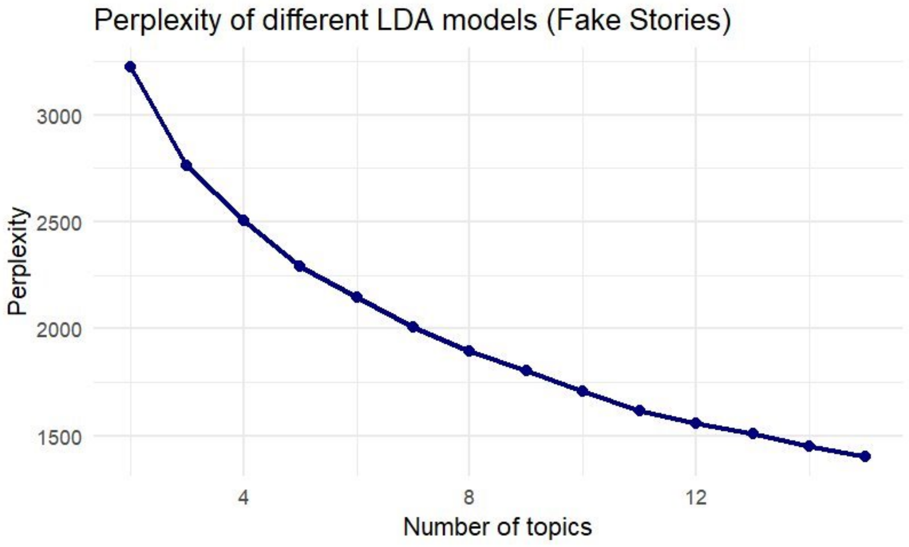
Sentiment Word Analysis Model for Story Data



Topic Modelling

Figure 11

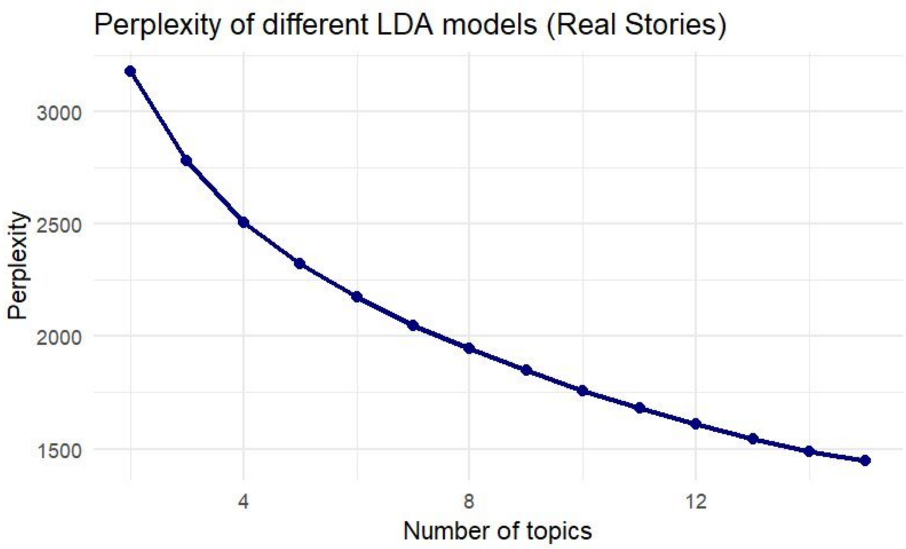
Perplexity of the LDA Models for Fake Release Data



In Figure 11, we see a visualization of the perplexity of sources that were deemed fake news. There is a clear downward trend in the perplexity levels of the LDA models as the number of topics increases. Thus, the coherency of the model increases as the number of topics does, meaning the model is capturing the structure of the fake news release data.

Figure 12

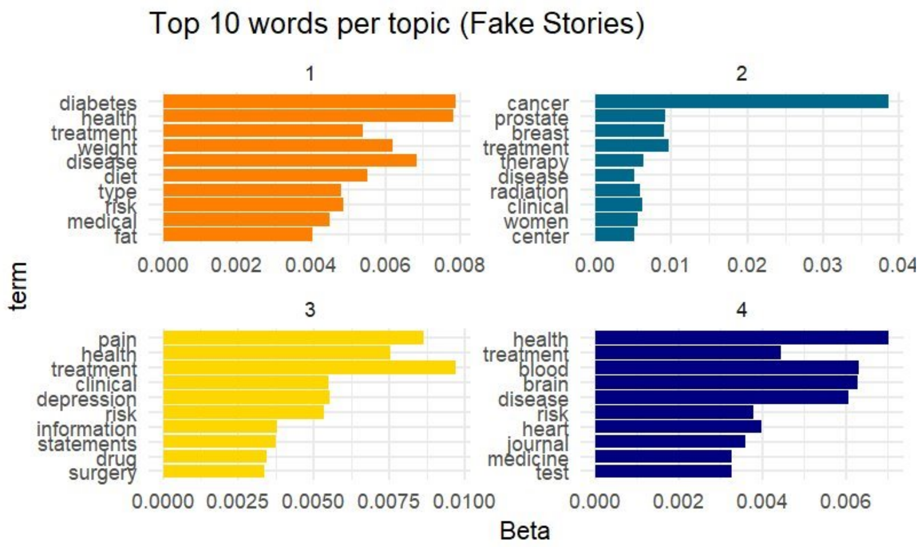
Perplexity of the LDA Models for the Real Release Data



In Figure 12, we see a similar plot to Figure 11, a perplexity model with a downward trend. While it is slightly different, the general trends are consistent with one another. Just as with our model concerning fake news, we can determine that the coherency increases as the number of topics increases. We can be sure that the structure of the real news release data is being captured as well.

Figure 13

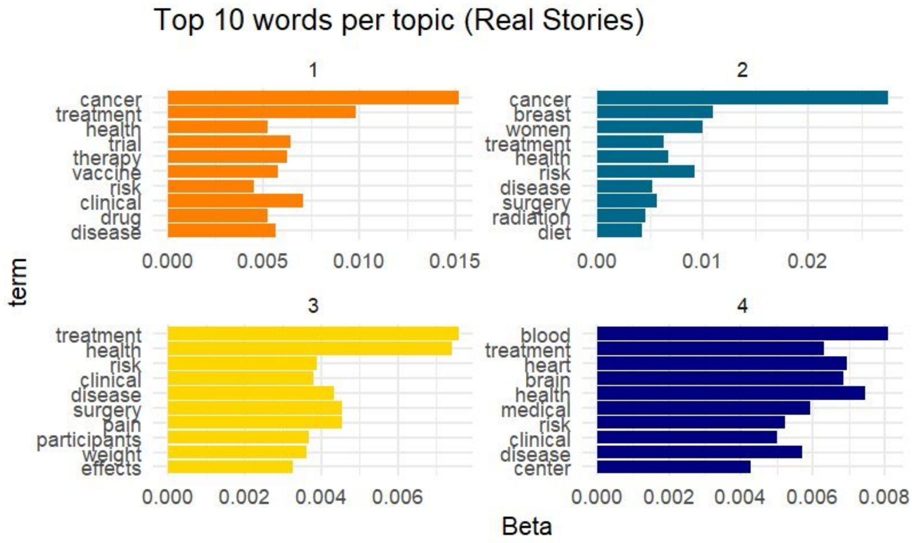
Top 10 Words per Topic Within the Fake Release Data



In Figure 13, we have four bar graphs for four different topics within the portion of the release data that was deemed fake news. When observing the various top words, we can see common words, such as ‘treatment’, and ‘risk’. When taking such trends into account, we can postulate that many of the fake articles within our dataset may be aimed toward people looking for home remedies to whatever ails them or are curious about the risks of certain medicines or procedures. Not only that, but we can glean that matters such as diabetes, cancer, and general pain commonly appear, meaning that fake news about those highly sensitive and life-altering health matters is especially persistent within our dataset.

Figure 14

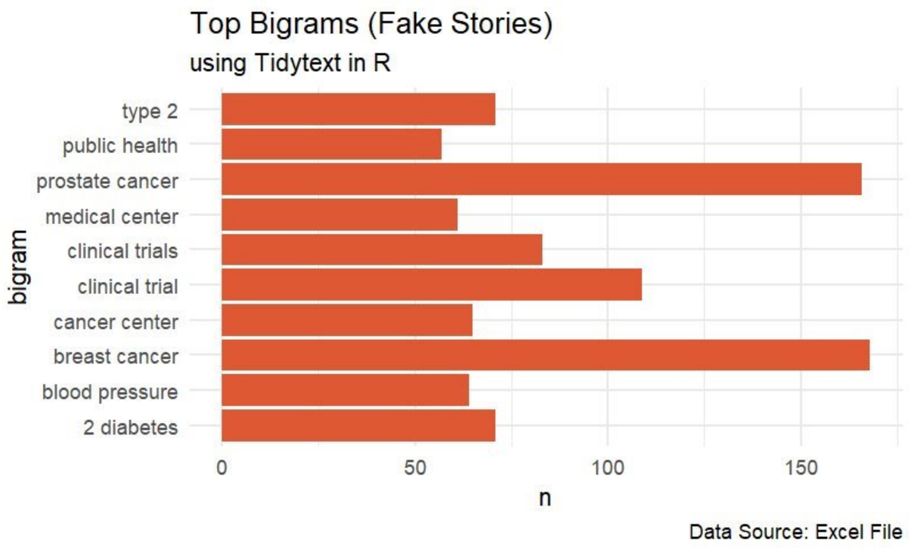
Top 10 Words per Topic Within the Real Release Data



In Figure 14, we have another set of bar graphs detailing the top ten words per topic, this time for release data that has been deduced to be real. Similarly to the fake news data, ‘treatment’, ‘risk’, and ‘health’ are all prominent words that frequently occur in all four topics, suggesting that even legitimate sources often discuss such topics, albeit without any sham information. ‘Cancer’ is a similarly often discussed subject, but ‘diabetes’ is notably missing, and ‘pain’ is not nearly as prominent. Instead, ‘blood’ is more prevalent, reflecting a difference in the priorities and subject prominence between real and fake news. We can speculate that perhaps health matters about blood may not appear as often in fake news due to it not being as much of a hot-button topic as diabetes or cancer, or as universal as pain.

Figure 15

Top Bigrams for Fake Release Data



In Figure 15, we see the top bigrams, the most prominent pairing of words, found within the fake news release data. Based on the results, reflecting what was found in Figure 13, ‘cancer’ is prominently featured within the bigrams, revealing they most often specifically referred to ‘prostate cancer’ and ‘breast cancer’. ‘Clinic trials’ is also heavily featured, though not nearly as prominently forms of cancer. This allows us to understand the general trends from Figure 13 in a greater context, revealing that fake news sources in our dataset focused prominently on breast and prostate cancer, both among the most common forms of cancer.

Figure 16

Top Bigram for Real Release Data

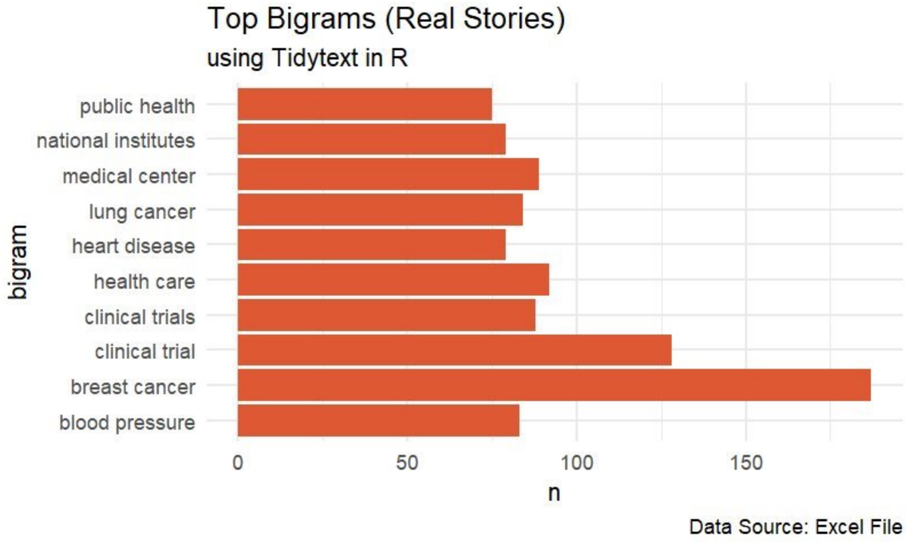
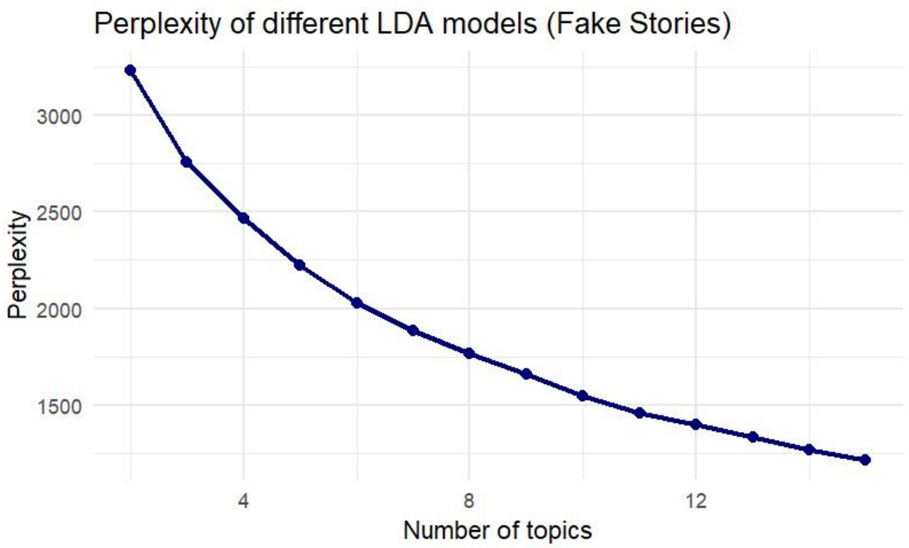


Figure 16 gives us bigrams of real news within the release dataset, offering us a greater context for the results found in Figure 4. Similar to Figure 15, ‘breast cancer’ is very prominent, as well as ‘clinical trials’, and there is a very noticeable lack of any mention of anything pertaining to diabetes. Real news within the release dataset focuses heavily on breast cancer and has a heavier emphasis on clinical trials compared to the fake news dataset.

Figure 17

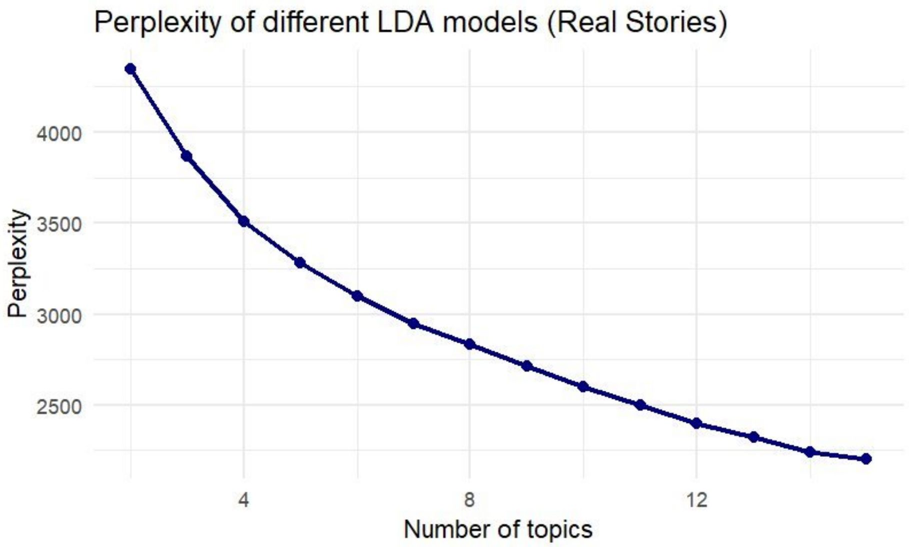
Perplexity of LDA Models for Fake Story Data



In Figure 17, we see a visualization of the perplexity of sources that were deemed fake news within the story dataset. There is a clear downward trend in the perplexity levels of the LDA models as the number of topics increases. Thus, the coherency of the model increases as the number of topics does, meaning the model captures the structure of the fake news story data.

Figure 18

Perplexity of LDA Models for Real Story Data

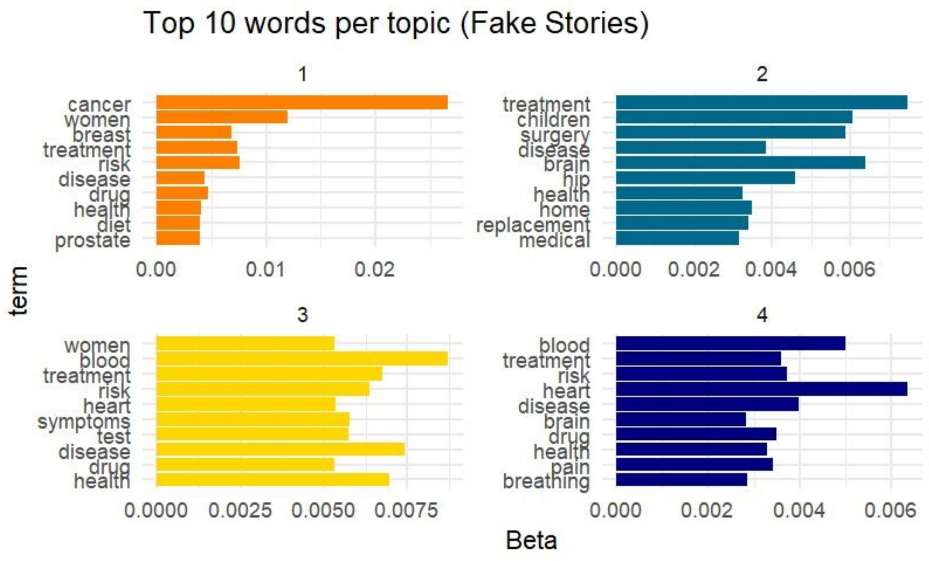


In Figure 18, we see a similar plot to Figure 17, a perplexity model with a downward trend.

While it is slightly different, the general trends are consistent with one another. Just as with our model concerning fake news, we can determine that the coherency increases as the number of topics increases. We can ensure the underlying structure of the real news story data is being captured too.

Figure 19

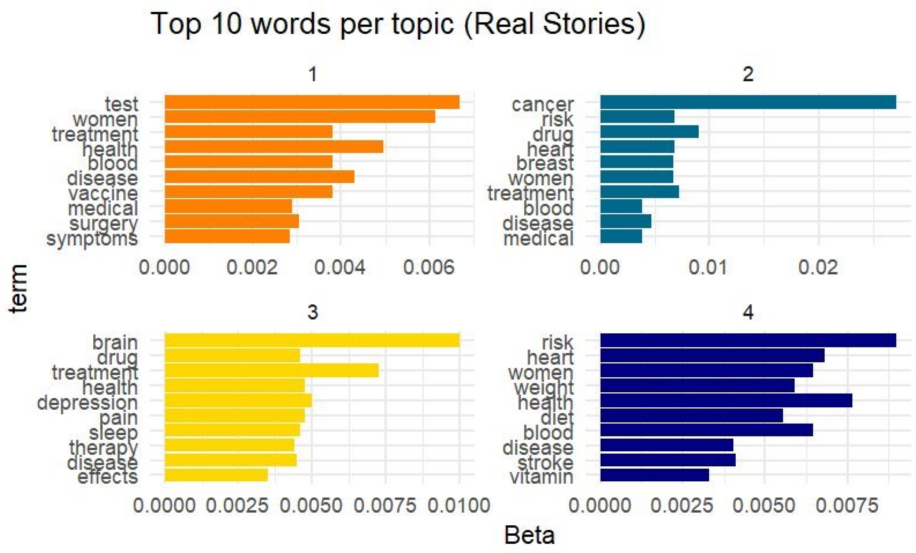
Top 10 Words per Topic for Fake Story Data



In Figure 19, we see four bar graphs detailing the top ten words per topic within the stories that were deemed fake news within the story dataset. We see that certain words are featured in all four graphs, such as ‘treatment’ and ‘health’. The most prominent words in each graph, in order, are ‘cancer’, ‘treatment’, ‘blood’, and ‘disease’. Based on the most prominent words, we can determine that matters of the blood and heart are quite subject to having misinformation spread about them, as well as cancer, similar to what we saw in Figure 13, and treatments for ailments or conditions.

Figure 20

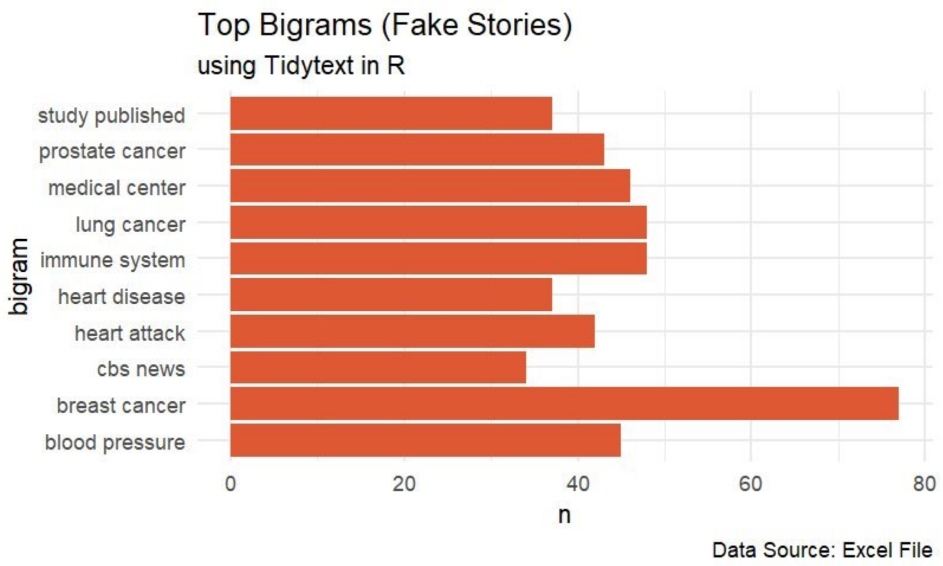
Top 10 Words per Topic for Real Story Data



In Figure 20, we can see 4 bar graphs that detail the top ten words for all four topics. We see that the most prominent word for each topic is, in order, ‘test’, ‘cancer’, ‘brain’, and ‘risk’. The word ‘disease’ is featured in all four topics, though to varying degrees of prominence. Based on this, we can see how real news and fake news observations diverge; while they both heavily discuss cancer, real stories tend to focus more on health care related tests, risks that presumably originate from procedures, and matters concerning the brain.

Figure 21

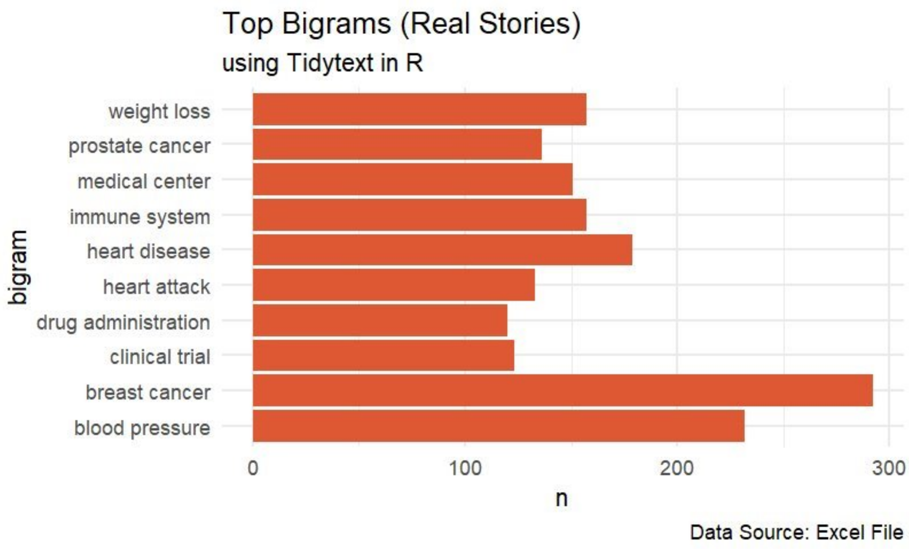
Top Bigrams for Fake Story Data



In Figure 21, we have a set of bigrams between the most prominent pairing of words with fake story data. ’Breast cancer’ is by a fair margin the most frequent pairing of words, with ’lung cancer’ and ’immune system ’being tied for the second most frequent. When cross-referenced with Figure 19, this development corroborates the significant value of cancer in topic 1. Interestingly, while it is not a particularly common pairing of words relative to the other bigrams, ’CBS News' is explicitly mentioned, perhaps indicating a prominent source of the fake news.

Figure 22

Top Bigrams for Real Story Data



In Figure 22, like Figure 21, ‘breast cancer’ is the most frequent pair of words, but this set of bigrams has a closer second-place pair in ‘blood pressure’. When cross-referencing Figure 20, we can see the pattern of cancer’s prominence repeating, seeing that it primarily refers to breast cancer specifically. While ‘blood pressure’ was also featured in Figure 21, it was not to the same degree as seen here, implying a greater emphasis on blood pressure in real news within our dataset.

Discussion of Results

The results of our analysis provide mixed support for both our hypotheses. Here we discuss our findings and their implications.

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

Our results provide some support for this hypothesis. Word frequency analysis and term frequency results found little support for overlapping common words between credible and less credible health publications leading us to conclude that surface-level word usage may appear fake and real health information. This finding reinforces the argument that identification of health misinformation, especially on the internet, can be a big challenge especially where people are in danger of consuming bogus and potentially harmful information on treatments. However, word ratio analysis and bing sentiment analysis found some evidence to support our hypothesis that health misinformation exhibits distinct linguistic patterns, with certain words recurring more frequently than others. Word ratio analysis found that some words like ‘cancer’ and ‘therapy’ and more specific to less credible health news. Again, *bing* sentiment analysis found that fake health publications tend to have a negative tone while more credible ones have an abundance of positive words. In general, our analysis of linguistic pattern suggests word usage may be insufficient in identifying health misinformation and analysis of thematic and narrative structure provides a more nuanced understanding of the mechanisms driving health misinformation.

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

The findings of our research provide support to this conclusion. When comparing and contrasting the top ten models and the bigram models between credible and non-credible sources, there were certain topics that had very commonly appeared, often in multiple different topics within the same model, for example, ‘cancer’, ‘blood’, ‘heart’, and ‘treatment’. So while we can safely say that misinformation had a high tendency to focus on similar topics, we can go one set further in our analysis and comment on how the topic modeling reinforces the notion that distinguishing between real and fake information at a glance, and especially for the average person who would not know what sort of information or clues to keep an eye out for. Overall, while the hypothesis was supported, it still does not provide a reliable solution that can be practically applied in order to address the problem of identifying health misinformation.

Overall, our findings suggest that the structure of health misinformation is nuanced and extremely complex. There are some differences between fake and real health news that we found and have pointed out, but these differences may not be consistent. Again, with the rapidly changing nature of the internet how health misinformation is presented may have changed dramatically.

Strengths and Limitations of the Study

For this study, we only had access to data for health stories and releases published before 2018. This is both a strength and a weakness of our study. By excluding documents from the Covid-19 period we get results that are more generalizable and less influenced by the pandemic-specific threads. However, not including more current publications provides a more conservative test for our hypothesis and we miss out on analyzing more recent patterns in health misinformation. Another strength of our study is the use of data from news media. In the United States, news media is still a significant source of information especially about new and trending matters. Nevertheless, we acknowledge that it is not the only or major source of health information. Thus, in terms of sampling, the study could be strengthened by using other sources of data such as social media channels including YouTube.

Works Cited

Goodyear, M. P. (2022). INHERENT POWERS AND THE LIMITS OF PUBLIC HEALTH FAKE NEWS. St. John's Law Review, 95(2), 319–378.

Hayawi, K., Shahriar, S., Serhani, M. A., Taleb, I., & Mathew, S. S. (2022). ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. Public health, 203, 23–30. https://doi.org/10.1016/j.puhe.2021.11.022

Huangfu, L., Mo, Y., Zhang, P., Zeng, D. D., & He, S. (2022). COVID-19 Vaccine Tweets After Vaccine Rollout: Sentiment-Based Topic Modeling. Journal of Medical Internet Research, 24(2), e31726–e31726. https://doi.org/10.2196/31726

Lyu, J. C., Han, E. L., & Luli, G. K. (2021). COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. Journal of Medical Internet Research, 23(6), e24435-. https://doi.org/10.2196/24435

Montesi, M. (2021). Understanding fake news during the Covid-19 health crisis from the perspective of information behaviour: The case of Spain. Journal of Librarianship and Information Science, 53(3), 454–465. https://doi.org/10.1177/0961000620949653

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019, September 18). Systematic literature review on the spread of health-related misinformation on social media. Social Science & Medicine. https://www.sciencedirect.com/science/article/pii/S0277953619305465

Zhou, X., Mulay, A., Ferrara, E., & Zafarani, R. (2020, October). Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 3205-3212).