Detecting Health Misinformation: A Text Analysis Approach

Data Visualization Project (BSDS 6301)

Oluwafunmibi Omotayo Fasanya

Louisiana State University Health Science Center, New Orleans.

November 11, 2025

Outline

Introduction

- Social media has accelerated the spread of misinformation, often faster than the truth.
- Misinformation in public health poses serious threat, which could lead to:
  - Avoiding vaccination or medical care
  - Using unverified or harmful remedies
  - Distrusting health authorities
- The World Economic Forum (2013) highlighted digital misinformation as a global threat.
- COVID-19 saw widespread misinformation, fueling vaccine hesitancy (Rodrigues et al., 2024).
- **Study goal:** Identify keywords and characteristics that distinguish fake vs. real COVID-19 headlines.

Methods: Dataset Overview

- The dataset used for this project is the COVID-19 Healthcare Misinformation Dataset (CoAID).
- It contains a collection of fake and real COVID-19 news gathered from:
  - Websites and verified news outlets
  - Social media platforms (Twitter, Facebook, etc.)
- Overall dataset summary:
  - **4,251** news articles
  - **296,000** user engagements
  - **926** social media posts
  - **Ground-truth labels** for all content
- This project focuses on 193 news titles (27 fake, 166 real) published on January 5, 2020.

Glimpse of the Dataset

		fact_check_url	news_url	title
1				
2	100000	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Spraying chlorine or alcohol on the skin kills viruses in the body"
3	100001	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Only older adults and young people are at risk"
4	100002	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Children cannot get COVID-19"
5	100003	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"COVID-19 is just like the flu"
6	100004	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Everyone with COVID-19 dies"
7	100005	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Cats and dogs spread coronavirus"
8	100006	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Face masks always protect against coronavirus"
9	100007	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Hand dryers kill coronavirus"
10	100008	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"SARS-CoV-2 is just a mutated form of the common cold"
11	100009	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"You have to be with someone for 10 minutes to catch the virus"
12	100010	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Rinsing the nose with saline protects against coronavirus"
13	100011	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"You can protect yourself by gargling bleach"
14	100012	medicalnewstoday.com	https://www.medicalnewstoday.com/articles/coronavirus-myths-explored	"Antibiotics kill coronavirus"

title
"How large does a meeting or event need to be in order to be a “mass gathering”?"
"Does WHO recommend that all international mass gatherings be cancelled because of COVID-19?"
"What factors should organizers and health authorities look at when assessing whether the risks are acceptable or not?"
"What if my organization does not have the expertise to assess the risks COVID-19 poses for our planned mass gathering?"
"If we go ahead with an international mass gathering, what can we do to reduce the risk of participants catching COVID-19?"
"Where can I find more advice on assessing and managing health risks around international mass gatherings?"
"What should be the criteria for excluding an athlete or other accredited participant from competing?"
"Should event organizers arrange screening at venues beyond national government requirements for point of entry (PoE)?"
"Should event organizers provide COVID-19 testing?"
"Are there additional safeguards event organizers can implement or recommend to athletes/officials/visitors in the context of COVID-19?"
"What are the risks arising from public transport to the venue(s)?"

Methods: Data Processing and Feature Engineering

- **Data Integration:**
  - Combined verified fake and real datasets into a single dataset.
- **Engineered Features:**
  - **Claim Length:** Total number of characters in the title.
  - **Word Count:** Number of words in each title.
  - **Average Word Length:** Claim length divided by word count — measures lexical complexity.
- These metrics provide insights into the structure and complexity of fake vs. real claims.

Methods: Sentiment Analysis

Overview

- Sentiment analysis was applied to capture emotional tone and opinion strength of each claim title.
- Two Python libraries were used for the sentiment analysis: `TextBlob` and `VADER`.

TextBlob

- A lexicon- and rule-based tool that uses a predefined dictionary of words with polarity scores.
- Extracts two key linguistic features:
  - **Polarity:** Measures emotional tone from -1(negative) to $+1$(positive).
  - **Subjectivity:** Measures degree of personal opinion, 0 (objective) to 1 (subjective).
- Captures general emotional orientation but less sensitive to punctuation or capitalization.

Methods: Sentiment Analysis (Cont'd)

VADER (Valence Aware Dictionary and Sentiment Reasoner)

- **Valence:** measures the emotional quality of each word.
- Designed for short, social-media-like texts — effective for claims and headlines.
- Computes:
  - **Positive, Negative, Neutral:** Proportions of sentiment within text.
  - **Compound:** Weighted normalized score from -1 (most negative) to $+1$ (most positive).
- Considers intensifiers, negations, punctuation (e.g., "!!!"), and capitalization for context-aware interpretation.
- Classification thresholds: Positive (≥ 0.05), Neutral (-0.05 to 0.05), Negative (≤ -0.05).

Methods: Certainty and Punctuation Analysis

Certainty Language Detection

- Created a custom dictionary of assertive words such as: *"absolutely"*, *"proven"*, *"guaranteed"*, *"undeniable"*, *"official"*.
- Computed a **certainty score** by counting occurrences of these terms in each title.
- A higher certainty score may indicate persuasive or misleading intent.

Punctuation Features

- Counted the number of **question marks** and **exclamation marks** in each title.
- Created binary indicators for the presence of these punctuations.
- Fake titles often use exaggerated punctuation to grab attention (e.g., "Can this cure COVID-19?!").

Visualization Techniques and Tools

Overview

- Visualizations were implemented in `Python` to explore linguistic, sentiment, and thematic patterns distinguishing fake and real COVID-19 claims.

Static Visualizations (Matplotlib & Seaborn)

- Violin/Box Plots
- Grouped Bar Charts
- Back-to-Back Histograms
- Word Frequency Bar Plots

Interactive Visualizations (Plotly)

- Scatter Plot

Lexical Visualization

- Word Cloud

Visual Results: Most Common Words

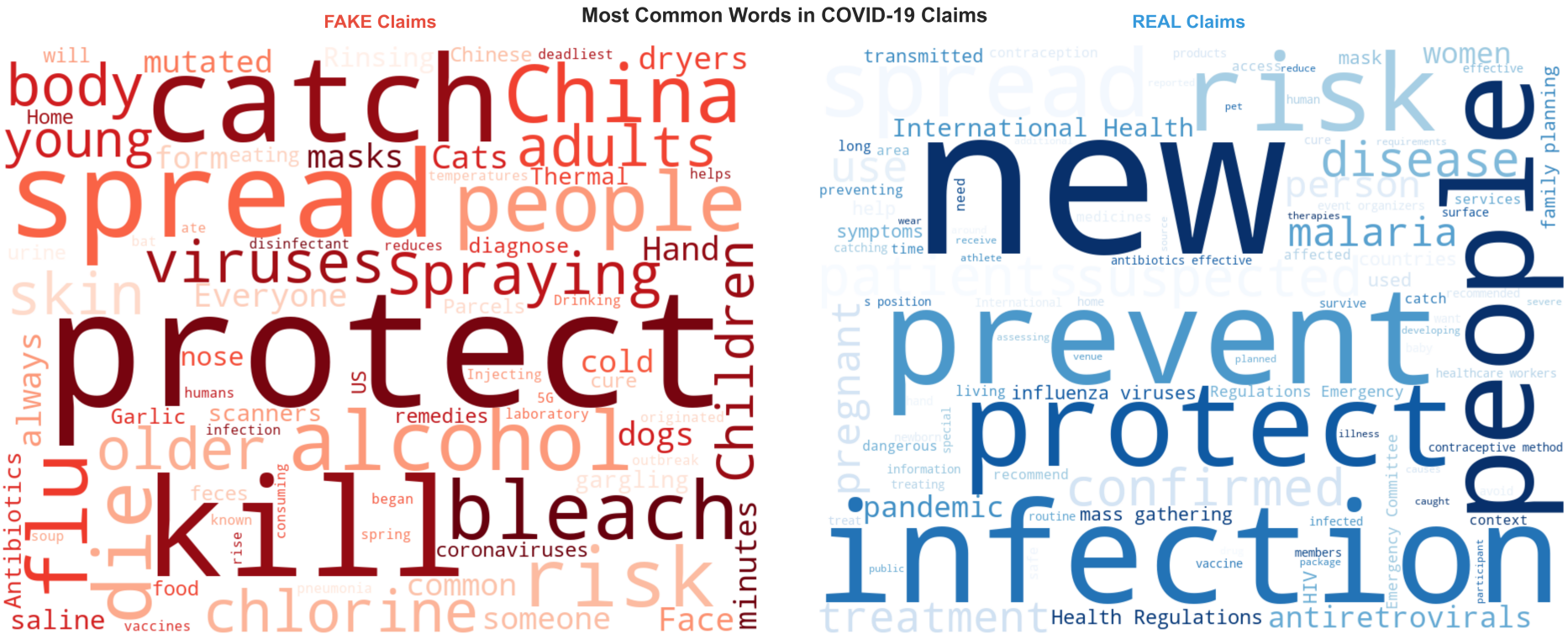

Figure 2a: Most Common Words in COVID-19 Headlines

Visual Results: Most Common Words (Unigram)

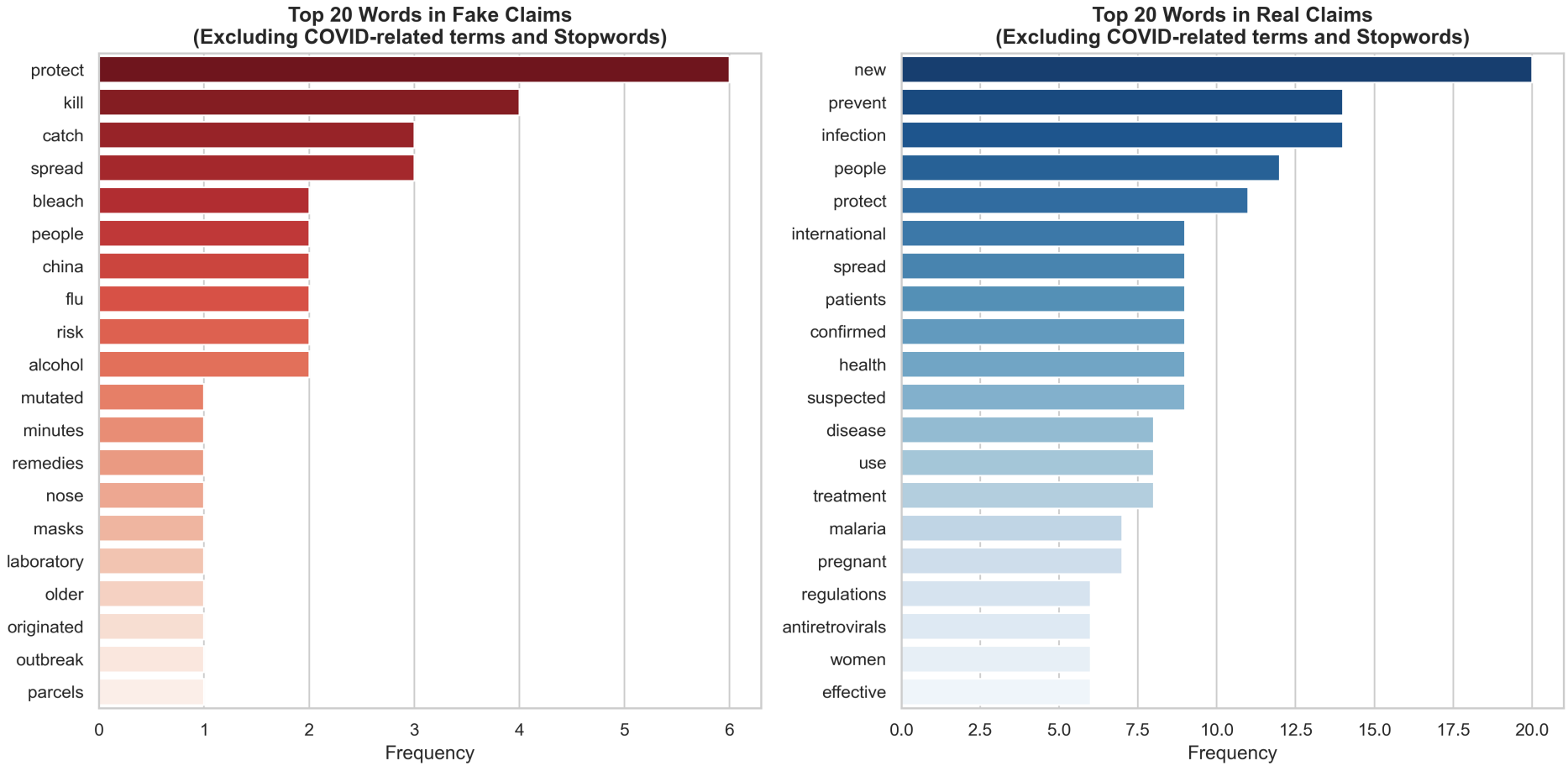

Figure 2b: 20 Most Common Words in COVID-19 Headlines

Visual Results: Most Common Words (Bigram)

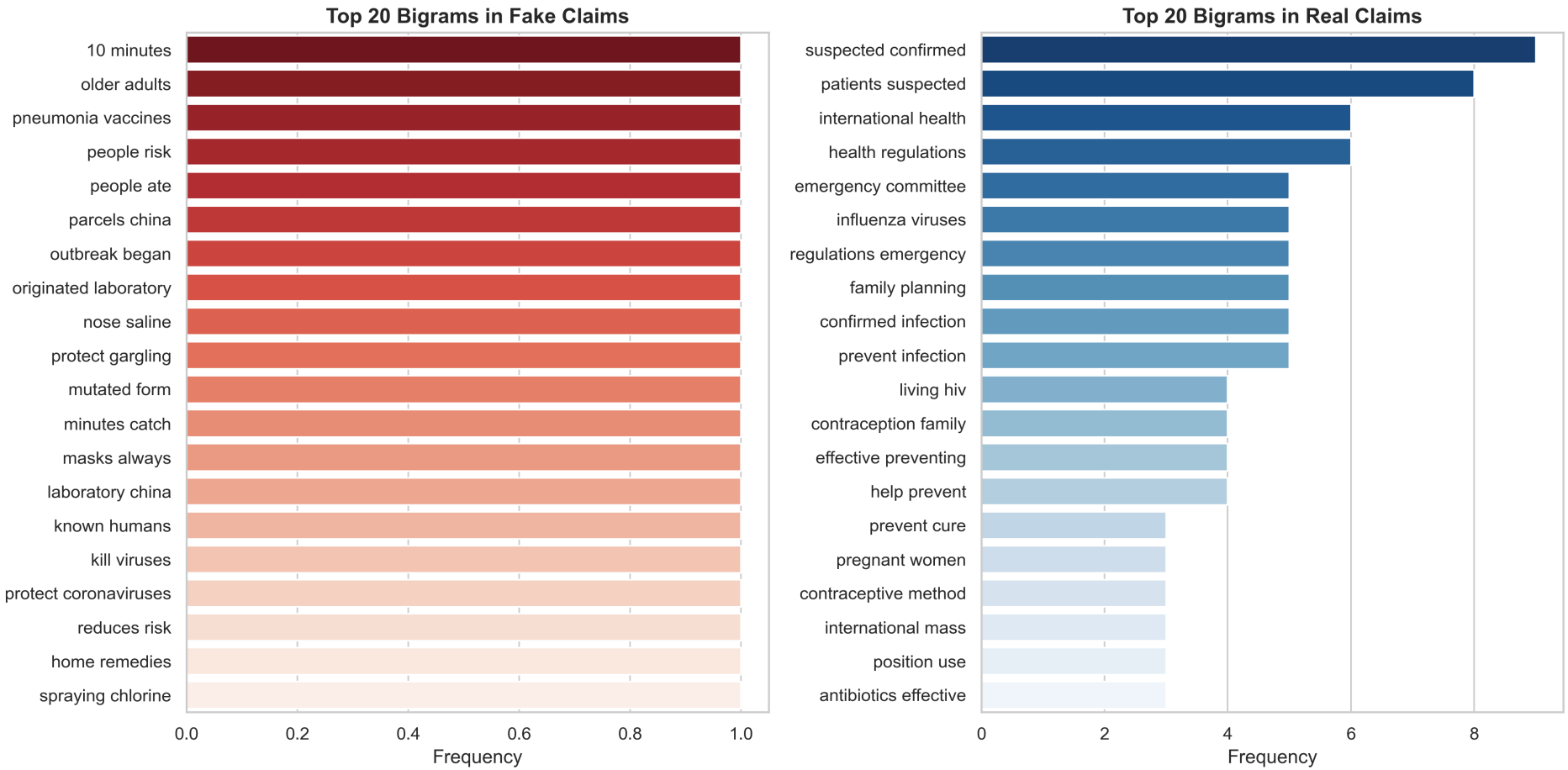

Figure 2c: 20 Most Common Bigram Words in COVID-19 Headlines

Visual Results: Most Common Words (Trigram)

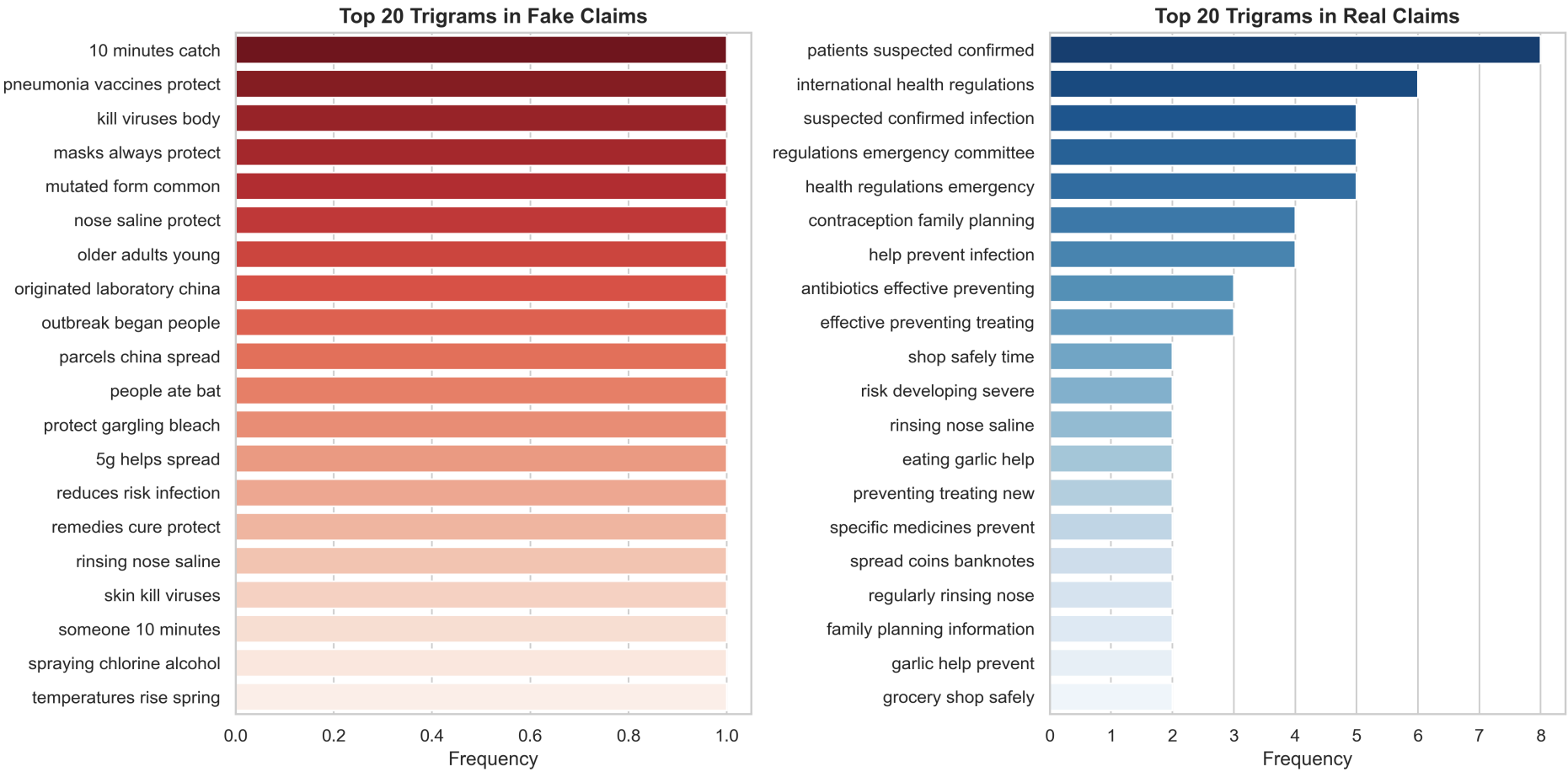

Figure 2d: 20 Most Common Trigram Words in COVID-19 Headlines

Visual Results: Overview Dashboard

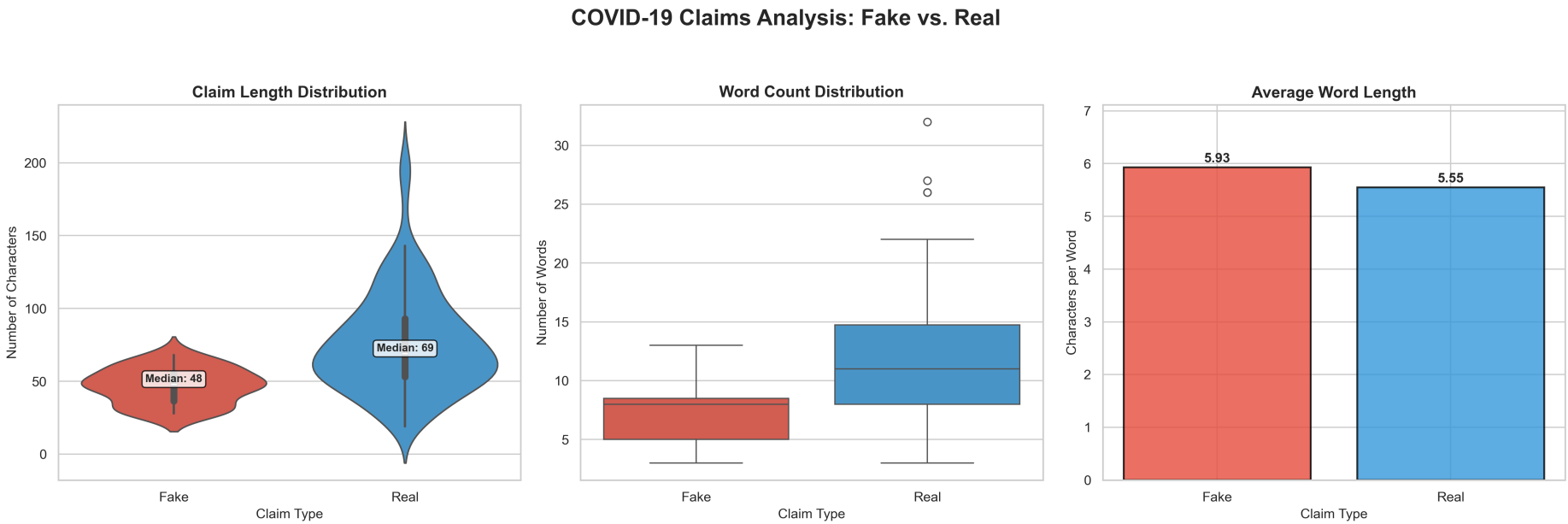

Figure 3: Overview Dashboard

Visual Results: Sentiment Analysis (Textblob)

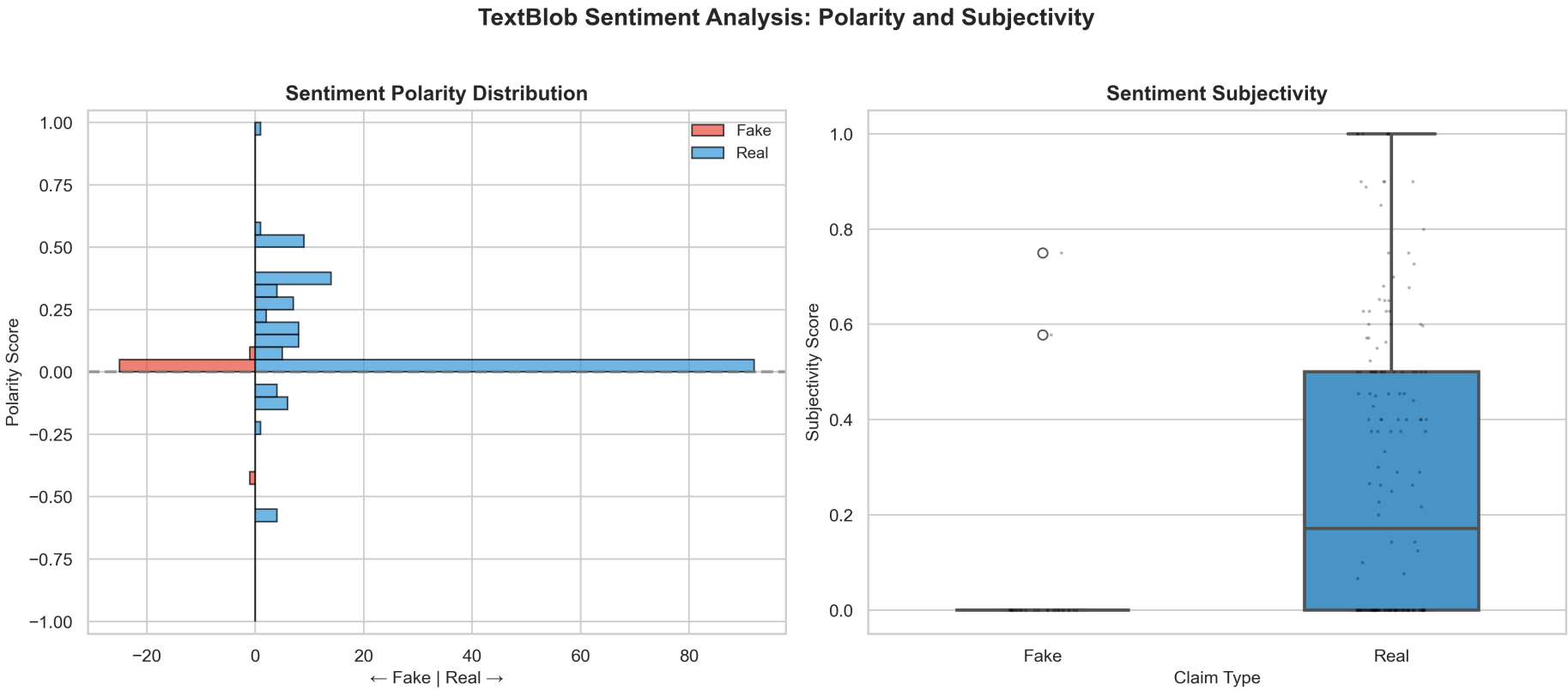

Figure 4: Polarity and Subjectivity

Visual Results: Sentiment Analysis (Textblob)

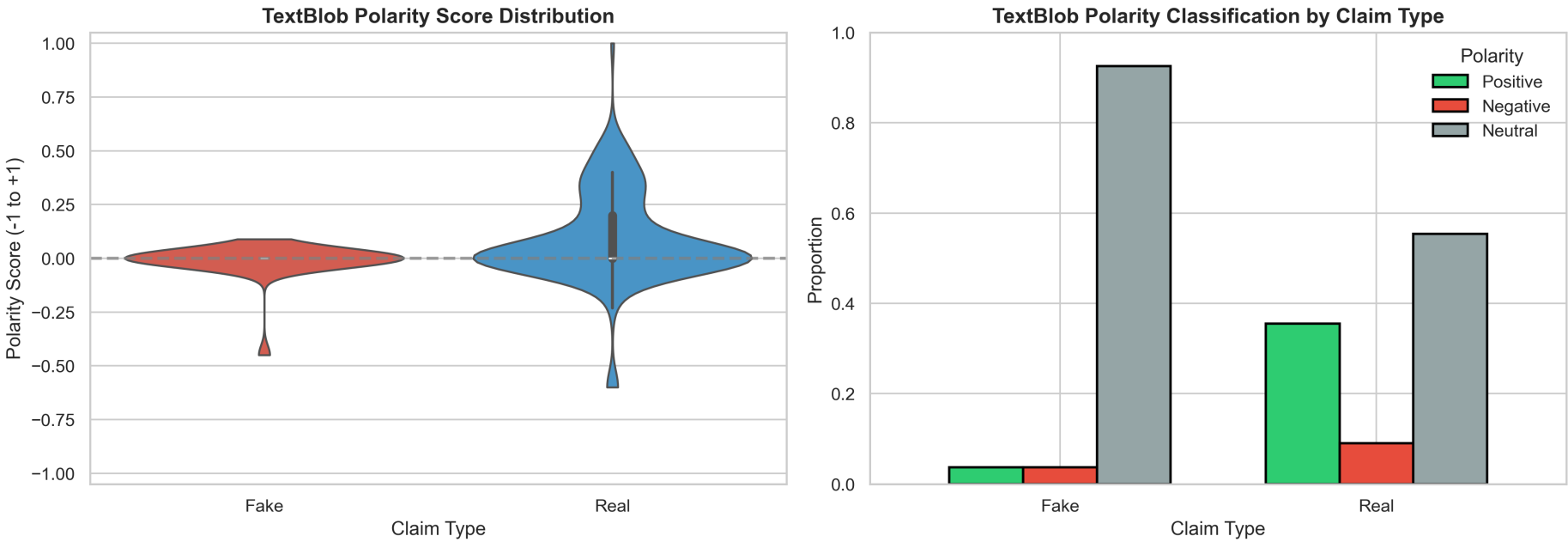

Figure 4: Polarity

Visual Results: Sentiment Analysis (Vader)

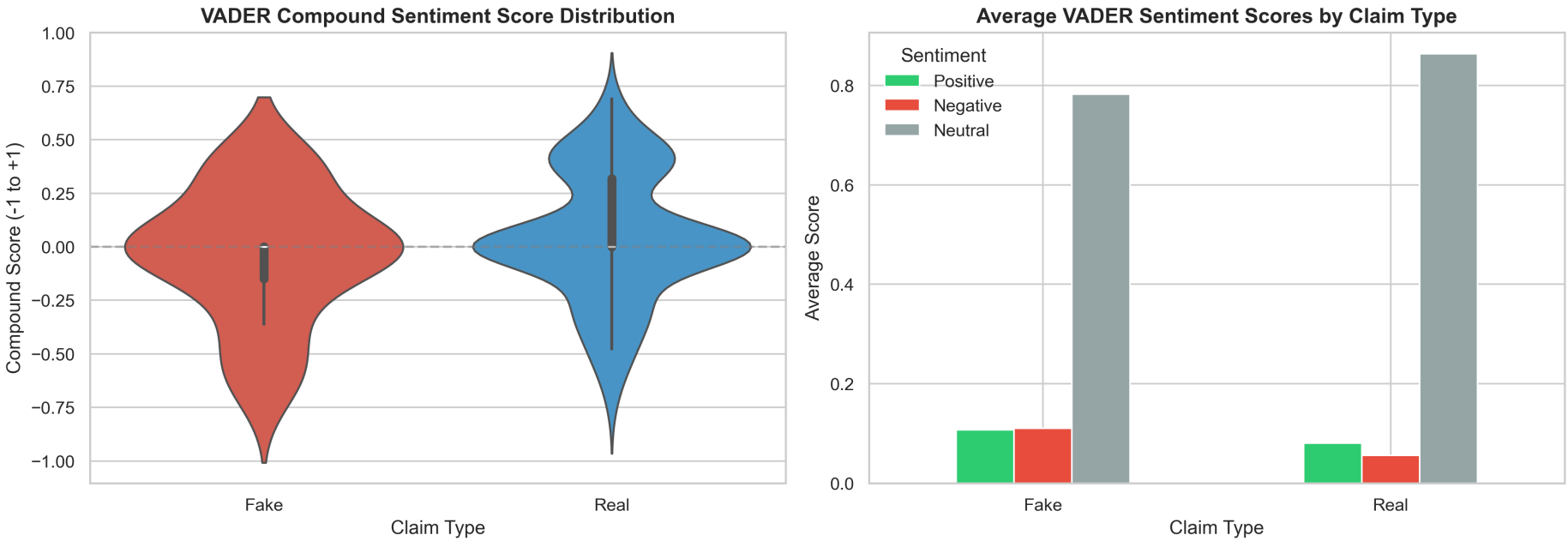

Figure 4: Sentiments in the Vader lexicon for each class

Visual Results: Sentiment Analysis (Textblob)

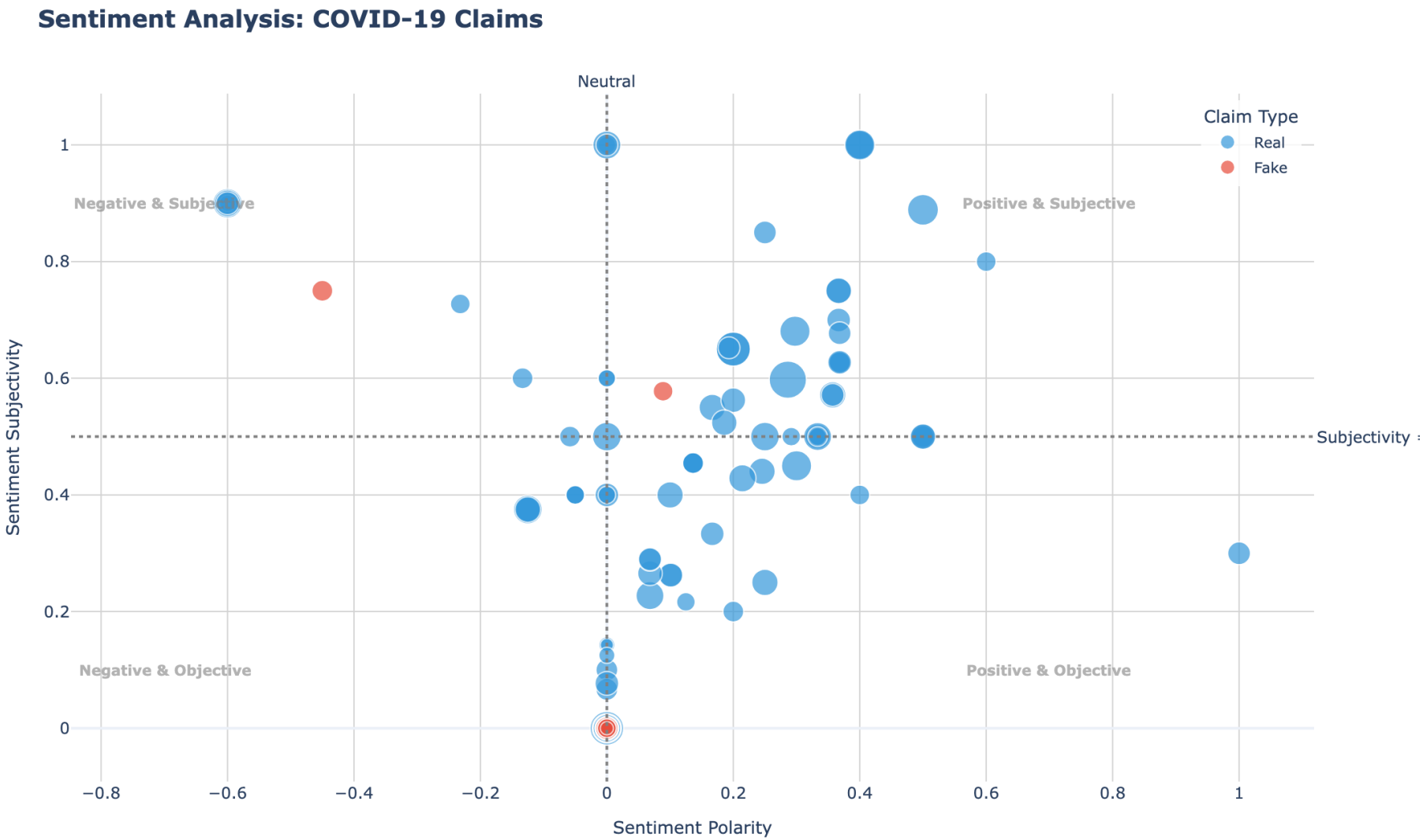

Figure 4: Sentiment Analysis

Visual Results: Feature Correlation Heatmap

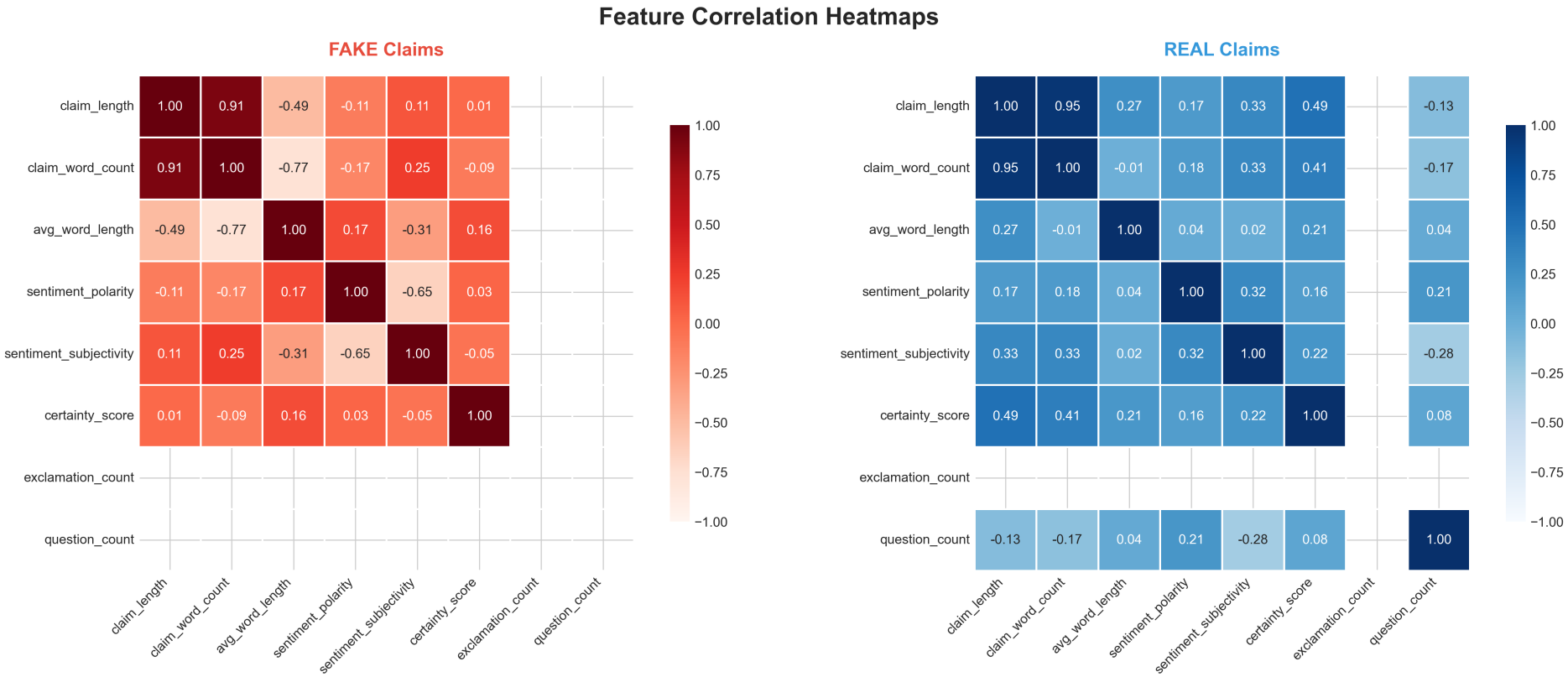

Figure 5: Sentiment Analysis

Future Work

- **Expand Dataset:** Incorporate more days beyond January 5, 2020, to improve model generalization and reliability.
- **Increase Sample Size:** Include additional fake and real claims to strengthen statistical power and reduce sampling bias.
- **Advanced Features:** Explore other linguistics such as modality, uncertainty, and emotional tone.
- **Interactive Visuals:** Develop a web-based dashboard for real-time exploration of fake versus real claims.

Questions?