King Saud University
College of Computer and Information Sciences
Department of Information Technology


IT 362 Course Project
Data Science
Semester-2, 1447H

# Sentiment analysis of opinions about flu vaccines

## Project Group #3



| **Student Name:** | **KSU ID:** |
|---|---|
| Jood Abdullah Alkhneen | 445201348 |
| Tala Abdullah Alqahtani | 445204552 |
| Shatha hamad mana | 445202291 |
| Rana Abdulrhman Alsalman | 445201204 |
| Lama Abdullah Almubarak | 445200338 |

**Supervised by:**
**Dr. Lama Al-Sudais**

# Introduction

Influenza vaccination is widely recognized as one of the most effective measures for preventing seasonal influenza and reducing its associated health risks. Despite its proven benefits, public attitudes toward flu vaccines remain diverse. While many individuals support vaccination and view it as essential for protecting personal and public health, others express concerns regarding vaccine safety, side effects, effectiveness, and trust in health authorities.

The widespread use of online platforms has enabled people to openly share their experiences, beliefs, and opinions about flu vaccination. These digital discussions generate large volumes of unstructured textual data that offer valuable insights into public perception and sentiment. Analyzing such data provides an opportunity to better understand vaccine acceptance, hesitancy, and the factors influencing public trust.

This project focuses on applying sentiment analysis techniques to examine public opinions about flu vaccines using data collected from multiple online sources. By investigating sentiment patterns and thematic trends, the study aims to provide a comprehensive understanding of how flu vaccination is perceived across different information channels.

**Research Question:**

What is the overall sentiment and thematic framing of flu vaccination across social media, regulatory reports, and news media sources?

# Literature Review

## Study 1: Examining the Negative Sentiments Related to Influenza Vaccination

**Problem addressed:**

This study investigated negative public sentiments toward influenza vaccination, focusing on vaccine hesitancy driven by misinformation and distrust in vaccination policies.

**Dataset used:**

Unstructured textual data from Twitter, consisting of 261,613 English tweets related to influenza vaccination collected between 2017 and 2022.

**Methods/models applied:**

BERT-based deep learning models were used for sentiment classification, and BERTopic was applied for topic modeling to identify major themes within negative sentiments.

**Key results:**

Negative sentiment toward influenza vaccination increased significantly after 2020. The dominant concerns involved distrust in government policies and the spread of misinformation.

---

## Study 2: Examining Public Messaging on Influenza Vaccine over Social Media

**Problem addressed:**

This study analyzed how organizations communicate about influenza vaccination on social media and assessed the scope and effectiveness of public health messaging.

**Dataset used:**

235,261 English tweets posted by organizational accounts between 2017 and 2023, collected using the Twitter API.

**Methods/models applied:**

Unsupervised deep learning topic modeling using BERTopic.

**Key results:**

Four main topics emerged: vaccination campaigns, vaccination during pregnancy, age recommendations, and vaccine importance during pregnancy. The study found limited diversity in public messaging content.

---

## Study 3: Examining Public Sentiments and Attitudes Toward Vaccination

**Problem addressed:**

This research explored public sentiments, vaccine hesitancy, and attitudes toward vaccination using large-scale social media discussions.

**Dataset used:**

2,944,530 English tweets collected from individual users between January and April 2021.

**Methods/models applied:**

CorEx topic modeling combined with VADER sentiment analysis. Preprocessing included duplicate removal and noise filtering.

**Key results:**

Six main thematic groups were identified, including vaccine policies, side effects, and hesitancy. Positive sentiment was generally higher, but hesitancy-related topics remained strongly negative.

---

## Study 4: Health Decision-Making Preferences and Influenza Vaccination

**Problem addressed:**

This study examined how individuals' health decision-making styles influence vaccine confidence and hesitancy.

**Dataset used:**

Survey data from nationally representative U.S. adult samples (1005 participants in 2016 and 1020 in 2018).

**Methods/models applied:**

OLS regression analysis based on health decision-making preference scales.

**Key results:**

Trust in science-based medicine and physician recommendations strongly increased vaccine confidence and reduced hesitancy.

---

## Study 5: Public Perceptions of Mandatory Influenza Vaccination Policies

**Problem addressed:**

This study explored public attitudes toward mandatory influenza vaccination policies for healthcare workers.

**Dataset used:**

1,163 online comments from Canadian news websites responding to vaccination-related articles.

**Methods/models applied:**

Qualitative thematic analysis with sentiment classification.

**Key results:**

Nearly half of commenters expressed negative sentiment toward flu vaccines, and most opposed mandatory vaccination policies due to concerns over personal freedom and vaccine safety.

---

# Comparison of Existing Studies

## Common Datasets and Features

Most studies relied on unstructured textual data from online platforms, particularly Twitter. Common features included opinion text, timestamps, and sentiment labels (positive, negative, neutral). The primary research focus across studies was understanding vaccine hesitancy and public attitudes.

---

## Differences in Modelling Approaches

Deep learning models such as BERT and BERTopic were used in Studies 1 and 2. Study 3 combined topic modeling with lexicon-based sentiment analysis (VADER). Study 4 applied statistical regression techniques using structured survey data, while Study 5 used qualitative thematic analysis. These approaches ranged from advanced machine learning to traditional statistical and qualitative methods.

---

## Strengths and Limitations

Social media-based studies benefit from large-scale real-time data but are limited by language constraints and demographic representation. Survey-based research provides behavioral insights but lacks dynamic public discourse. Qualitative studies offer detailed thematic understanding but involve smaller datasets. Overall, most studies focus on sentiment trends but lack integration across multiple data sources.

---

## What Is Missing or Limited in Previous Work

Although existing research provides valuable insights into public attitudes toward vaccination, several limitations remain. Many studies rely on single data sources and English-language content, reducing generalizability. Additionally, limited research integrates multiple perspectives such as regulatory data and media coverage. Some qualitative studies also use relatively small sample sizes, restricting broader conclusions.

---

## How This Project Builds Upon Existing Studies

This project focuses specifically on public opinions toward influenza vaccination and collects data from multiple sources, including online discussions, official health reports, and news media coverage, to provide a broader and more diverse perspective. By emphasizing influenza vaccines and utilizing varied data sources, this study aims to support a more comprehensive understanding of vaccine acceptance and hesitancy.

# Data Sources and Bias Evaluation

## Data Source

This project combines data from multiple publicly available APIs to obtain diverse perspectives on public opinions regarding flu vaccination. By integrating social media discussions, regulatory health reports, and news media coverage, the dataset aims to provide a more comprehensive understanding of how flu vaccines are perceived across different information channels.

---

### Source 1: YouTube Data API v3

The YouTube Data API was utilized to collect public comments related to flu vaccination. Relevant videos were identified using keyword-based searches such as “flu vaccine,” “flu shot,” and “influenza vaccine,” and associated comment threads were extracted.

- Number of observations: 3,896 comments
- Main features and data types:
  - comment_text (string – unstructured text)
  - publishedAt (datetime)
  - video_id (string)
- Data format: JSONL (unstructured textual data)

This source captures spontaneous public reactions and discussions surrounding flu vaccines.

---

## Source 2: openFDA API

The openFDA API was employed to retrieve reports related to adverse events following influenza vaccination from the FDA's pharmacovigilance system.

- Number of observations: approximately 1,000 reports
- Main features and data types:
  - report_id (string)
  - reaction (categorical/text)
  - patient_age (numeric)
  - report_date (datetime)
- Data format: Structured JSON records

This dataset provides an official regulatory perspective on vaccine-related outcomes rather than subjective opinions.

---

## Source 3: GNews API

The GNews API was used to collect news articles discussing influenza vaccination through keyword-based queries.

- Number of observations: 110 articles
- Main features and data types:
  - title (string)
  - description (string)
  - content (string – semi-structured text)
  - publishedAt (datetime)
  - source (string)
- Data format: Semi-structured JSON

This source reflects how flu vaccination is portrayed in media coverage.

---

# Bias Evaluation

Given the integration of multiple data sources, several forms of bias may influence the dataset.

## Representation Bias

The collected data may not equally represent all population groups. YouTube comments primarily reflect individuals who actively engage online, potentially overrepresenting younger or more digitally active users. openFDA reports tend to capture individuals who experienced and reported adverse events, which may emphasize negative outcomes. News articles are shaped by editorial priorities and may highlight controversial topics.

Consequently, individuals with limited internet access or those who do not report vaccine experiences may be underrepresented.

---

## Measurement Bias

Bias may arise from the manner in which data is produced and collected. Social media comments may include sarcasm, emotional language, or misinformation. openFDA reports depend on voluntary reporting and the accuracy of medical documentation. News articles reflect journalistic framing rather than neutral observation. In addition, keyword-based API searches may influence which records are retrieved.

---

## Historical Bias

Public discussions about vaccines are often shaped by previous health crises, misinformation campaigns, and societal trust in healthcare institutions. Adverse event reporting systems historically capture more severe cases, while media narratives may reflect longstanding public debates and inequalities in healthcare access.

As a result, the dataset may mirror existing societal attitudes rather than an unbiased distribution of opinions.

# Objectives

The main objective of this project is to analyze public opinions about flu vaccines using unstructured and semi-structured textual data collected from multiple online sources.

Specifically, this study aims to:

1. Identify the overall sentiment (positive, negative, and neutral) expressed toward flu vaccines.
2. Examine common concerns and themes associated with negative and positive opinions.
3. Compare sentiment patterns across different data sources, including social media, regulatory reports, and news media.
4. Explore changes in public sentiment over time.
5. Generate insights that can support public health awareness and communication strategies.

# Method

This project follows a structured data collection and analysis workflow to examine public sentiment toward flu vaccines using unstructured and semi-structured textual data obtained from multiple online sources.

First, data was collected using publicly available APIs, including the YouTube Data API, openFDA API, and GNews API. Keyword-based queries such as “flu vaccine,” “flu shot,” and “influenza vaccine” were used to retrieve relevant records. The collected data was stored in its raw format (JSON/JSONL) without any modifications to preserve its original structure.

Next, basic preprocessing steps will be applied in later phases of the project to prepare the textual data for sentiment analysis. These steps are expected to include removing duplicates, handling missing values, text normalization, tokenization, and noise removal.

Following preprocessing, sentiment analysis techniques will be employed to classify the collected text into positive, negative, or neutral categories. Both traditional machine learning approaches and deep learning models may be explored to compare performance.

In addition to sentiment classification, topic modeling methods will be used to identify common themes and concerns expressed in public discussions about flu vaccination. Temporal analysis may also be conducted to examine how sentiment trends change over time.

Finally, the results will be visualized and interpreted to provide insights into public attitudes, potential misinformation patterns, and overall perception of flu vaccines.

# Challenges

During Phase 1 of the project, several challenges were encountered in the data collection process. Initially, web scraping techniques were attempted on platforms such as Reddit and health review websites. However, these attempts were blocked by access restrictions and HTTP 403 errors, making automated scraping unreliable for large-scale data collection.

Additionally, some sources provided limited or inconsistent data volumes, which made it difficult to build a scalable and comprehensive dataset. As a result, the team shifted to using API-based data collection methods, which offered more stable and structured access to information.

Despite the advantages of APIs, further challenges arose, including configuring authentication systems (such as Google Cloud setup), handling pagination across large result sets, and managing free-tier rate limits imposed by certain services, particularly the daily request limits of the GNews API. These limitations required careful adjustment of query parameters and strategic selection of data sources to ensure sufficient raw data was collected without modification.

To mitigate these challenges in future stages, it is recommended to apply for extended API access tiers, utilize automated scheduling tools to distribute data requests over time, and explore official research access programs provided by data platforms. Moreover, maintaining multiple alternative data sources can reduce dependency on a single platform and improve overall data availability.

# References

1-Examining the Negative Sentiments Related to Influenza Vaccination from 2017 to 2022: An Unsupervised Deep Learning Analysis of 261,613 Twitter Posts - PMC

2-Examining Public Messaging on Influenza Vaccine over Social Media: Unsupervised Deep Learning of 235,261 Twitter Posts from 2017 to 2023 | MDPI

3- https://pmc.ncbi.nlm.nih.gov/articles/PMC9014796/

4-Understanding influenza vaccination attitudes and behaviors: An assessment of health decision-making preferences - ScienceDirect

5- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129993