COVID-19 Vaccine Sentiment Analysis

Levente Szabo & Ryan Cohen

Overview

1. Introduction: Problem Statement
2. Vaccine Analysis Pipeline
3. Data Modelling
4. Data Ingestion
5. Data Analysis
6. Data Visualization

Introduction

- The development of a vaccine that could end the global COVID-19 pandemic would require a large scale program to vaccinate a majority of the population.

- It is estimated that in order for COVID-19 herd immunity to be achieved, **82% of the population** would have to be immune to the virus.

- With multiple vaccines in rapid development we anticipate a variety of barriers regarding widespread vaccination.

- In order to gauge public perception regarding vaccination, we construct a **sentiment analysis system** that analyzes Twitter data in real time.

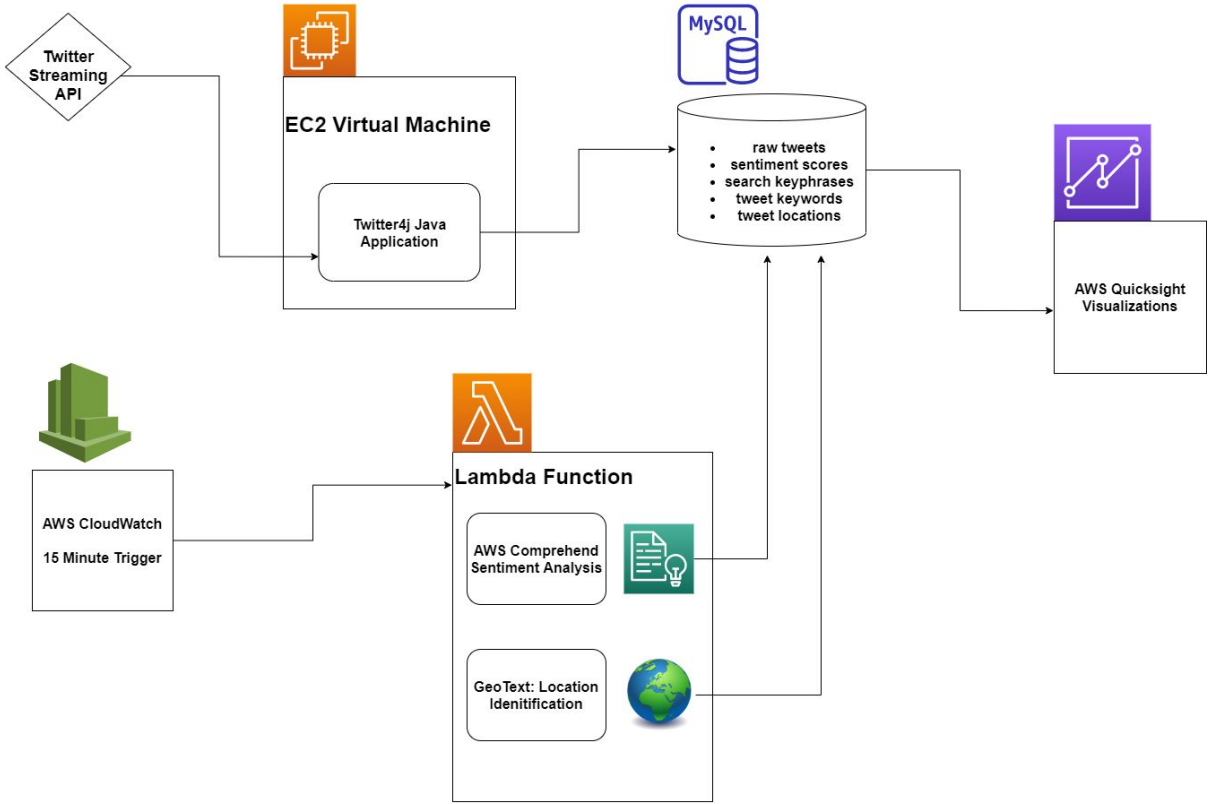
Introduction

- We aim to study factors that shape individuals perceptions of vaccination.
  - Examples: Fear, Conspiracy, Trust

- Additionally, we aim to study mentions of key opinion leaders.
  - Examples: Moderna, Dr. Fauci, Pfizer

- The analysis of tweets can allow for study of geographic and temporal differences in vaccine sentiment.
  - Sentiment of related tweets can be tracked over time.
  - US and world maps will be created using user-provided location descriptions.

Vaccine Analysis Pipeline



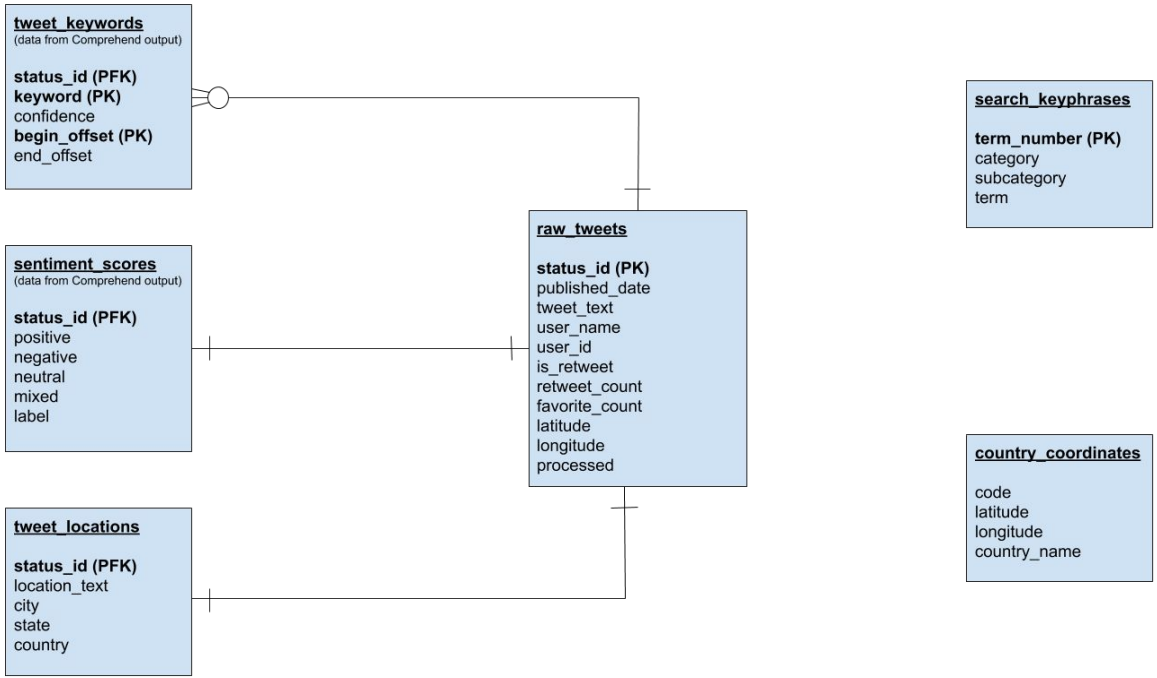
Vaccine Analysis Pipeline

1. **Ingestion:** Java application using Twitter4j deployed on an EC2 virtual machine. Tweets mentioning “vaccine” and other predetermined keywords are preprocessed and stored into a MySQL relational database.

2. **Analysis:** On regular time intervals a CloudWatch service initiates a serverless Lambda function to read unprocessed tweets from the MySQL database. Sentiment analysis and Keyword extraction are performed with Comprehend and location matching using the Python geotext package.

3. **Visualization:** Analytics regarding the breakdown of tweet sentiments, most common keywords and user locations is displayed as a dashboard using QuickSight. Data is constantly refreshed to provide a real-time user experience.

Database Modelling: Schema



Category Keywords & Keyphrases

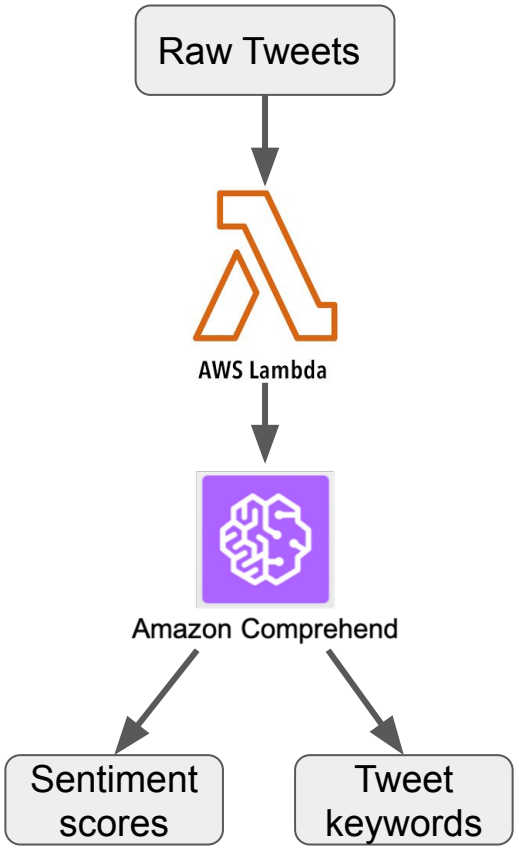
Sample Category	Sample Keywords & Keyphrases
Conspiracy	“microchip vaccine”, “5G”, “Gates”
Companies	“Pfizer”, “Johnson & Johnson”, “Oxford”
Hesitancy	“vaccine accessibility”, “vaccine profiteering”
Safety	“Vaccines kill”, “vaccines are poison”
Trust	“Experimenting on us”, “I am pro-vaccine”

Data Ingestion - Tweet Collection

- Twitter4j to interface with Twitter and JDBC to interface with MySQL
- To perform continuous tweet ingestion our application is run on an AWS Linux EC2 instance.
- To determine the set of keywords to filter on we initially query the search_keyphrases table in our database.
- Many tweets are in fact retweets of popular tweets. To avoid overloading our analysis system we handled these retweets by updating the *retweet_count* and *favorite_count* in our database.

Data Analysis - Sentiment & Keywords

1. Lambda function is triggered every 15 minutes by an AWS CloudWatch event.
2. Lambda function selects unprocessed raw tweet data from the database.
3. For each unprocessed raw tweet, the lambda function calls AWS Comprehend to
    a. Measure sentiment (positive/negative/neutral/mixed ratings and overall label)
    b. Extract keywords (key phrases and confidence scores)
4. Store results in sentiment and keyword tables.



Data Analysis - Location Extraction

1. City, state, and country variables initialized to None.
2. Using results identified by the geotext library, set the city and country variables.
3. If the country is “US” or unidentified:
    a. Split the location text on commas
    b. For each resulting token, check if it is present in predefined sets of US state names and two-letter state abbreviations
        i. If so, state is set to token, and country is set to “US”

Examples:

- “Evanston, IL” ⇒ city=“Evanston” state=“IL” country=“US”
- “Toronto” ⇒ city=“Toronto” state=None country=“CA”
- “Way north” ⇒ city=None state=None country=None

Data Visualization - QuickSight Dashboard

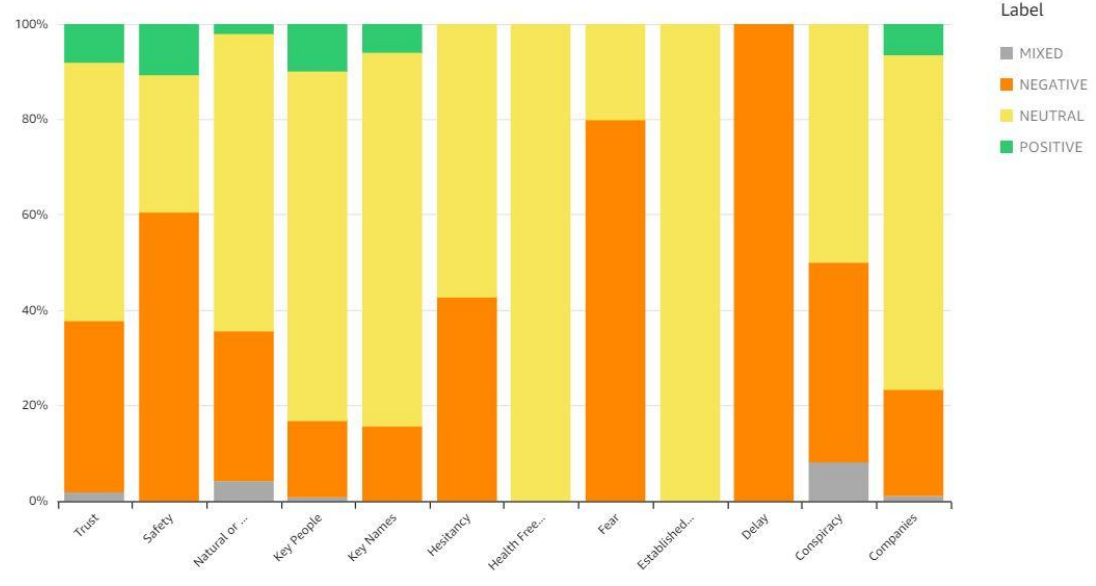
- Figures in a QuickSight dashboard reflecting insights from the collected tweets update on a daily basis.
- All figures other than the time series reflect insights from the past seven days of collected tweets.
- Dashboard consists of three sets of figures:
  1. Insights for tweets matching pre-specified category terms
  2. Sentiment and collected tweet count time series
  3. Geographic and AWS Comprehend keyword data

Data Visualization - Categories

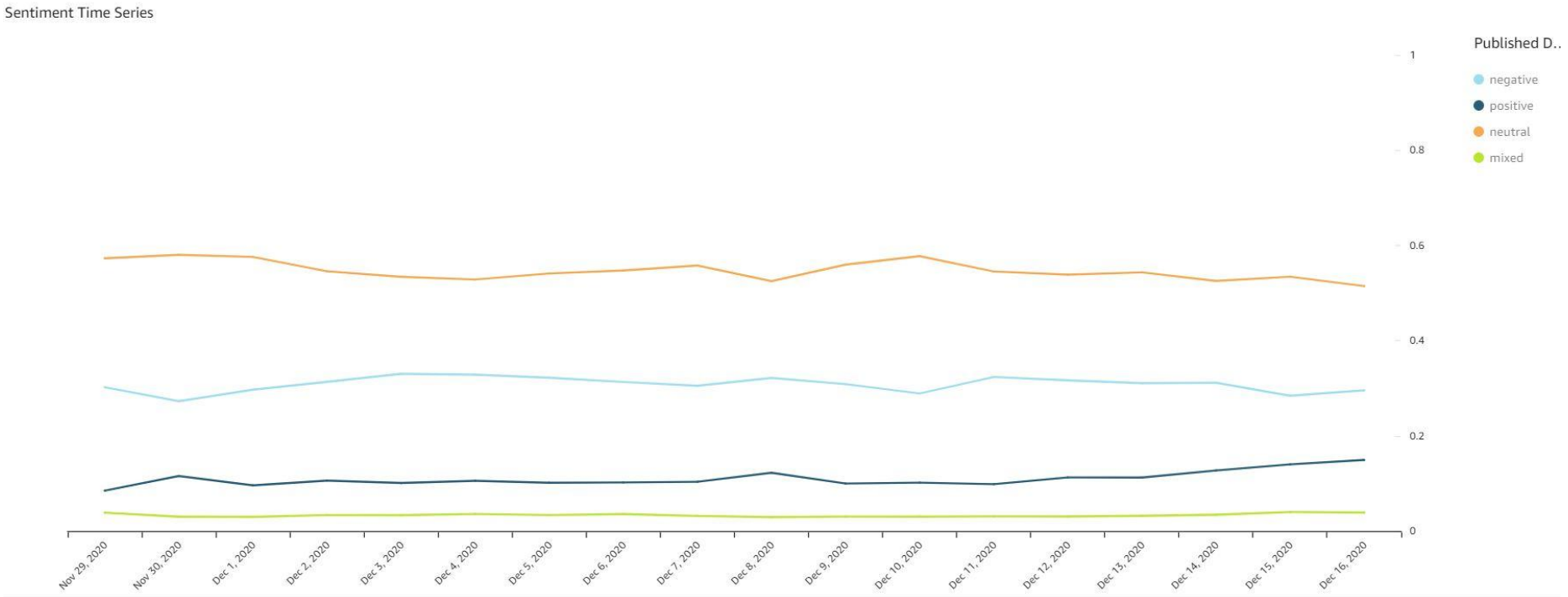
Number of matching tweets per category

Category	Count
Companies	35,918
Key People	399
Trust	114
Key Names	51
Natural or Alternatives	48
Safety	38
Conspiracy	12
Hesitancy	7
Fear	5
Established Sources	3
Delay	1
Health Freedom	1

Sentiment breakdown for tweets in each category



Data Visualization - Time Series



Data Visualization - Maps