CalPolyPomona

# DATA COLLECTION AND PREPROCESSING

*REPORT*

**GBA 6410: Social Media Analytics and Text Mining**

Instructor: Dr. Mehrdad Koohikamali

Ailien Dang

Cece Nguyen

Kiet Nguyen

Ngoc Nguyen (Alice)

Ngoc Nguyen (Kim)

Vy Nguyen

**Introduction**

This report details the process of collecting, preprocessing, and analyzing data obtained from the U.S. Department of Health and Human Services (HHS) Office for Civil Rights breach report portal. The goal is to prepare clean data suitable for analysis and to perform descriptive analysis on selected keywords.

**Source Data**

The source data was obtained from the U.S. Department of Health and Human Services Office for Civil Rights breach report portal, available at HHS OCR Breach Report.

**Data Collection Details**

- **Source**: U.S. Department of Health and Human Services Office for Civil Rights
- **Method**: The data is downloaded directly from the archive section of the source website.
- **Frequency**: All breaches reported within the last 24 months that are currently under investigation by the Office for Civil Rights.
- **Format**: The data is typically available in formats such as CSV, Excel, and JSON.

**Data Preprocessing and Cleaning Stages**

1. **Handling column name inconsistency**

Standardizes the column names of a DataFrame by replacing spaces and hyphens with underscores and converting all letters to lowercase.

2. **Data Type Conversion**

- The *breach_submission_date* column was converted to a datetime format, enabling time-based analyses and proper handling of date-related operations.
- ensures that the columns ‘name_of_covered_entity’, ‘type_of_breach’, ‘state’, and ‘web_description’ in the DataFrame data are all converted to string data type.

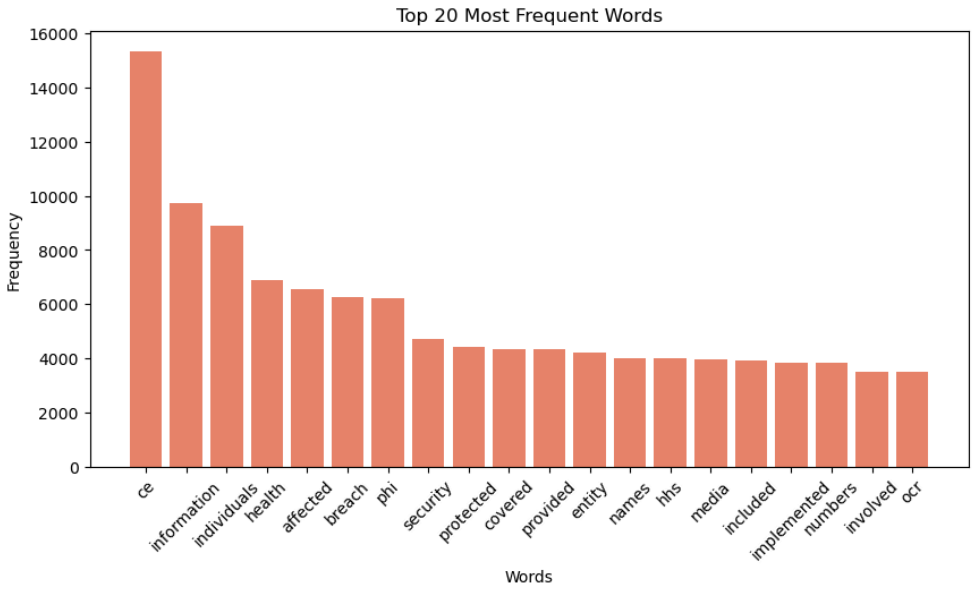
1. **Handling Missing Values**

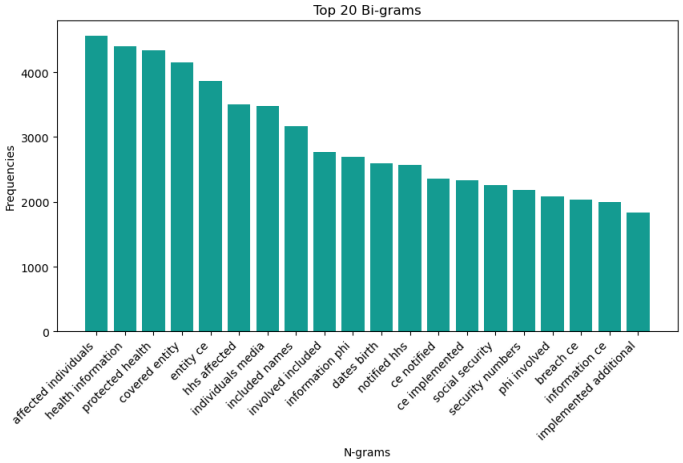
- *Proportion of Missing Values:* The proportion of missing values was calculated to assess the overall data quality (0.00012%)
- *Mode Imputation for Object Columns:* For categorical columns containing missing values, the mode (most frequent value) was used to fill in the gaps.
- *Forward Fill for Datetime Column:* Missing values in the *breach_submission_date* column were filled using the previous valid date (forward fill).
- Replaces any missing values in the *'individuals_affected'* column with the column's average value.
- Removes duplicate rows from the DataFrame

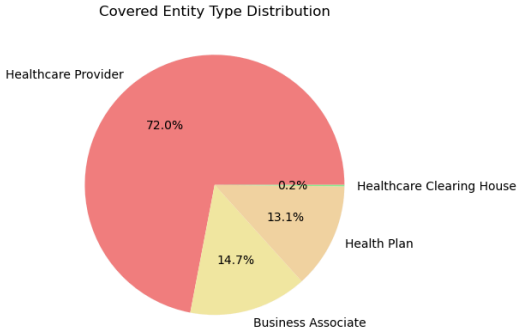
**Text Data Preparation**

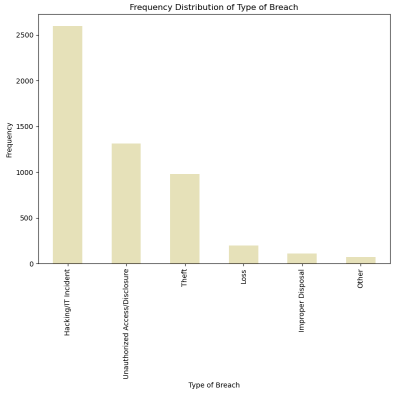
- A text corpus was built from the *web_description* column.
- All elements in the *web_description* column were ensured to be strings, facilitating uniform text processing and analysis.

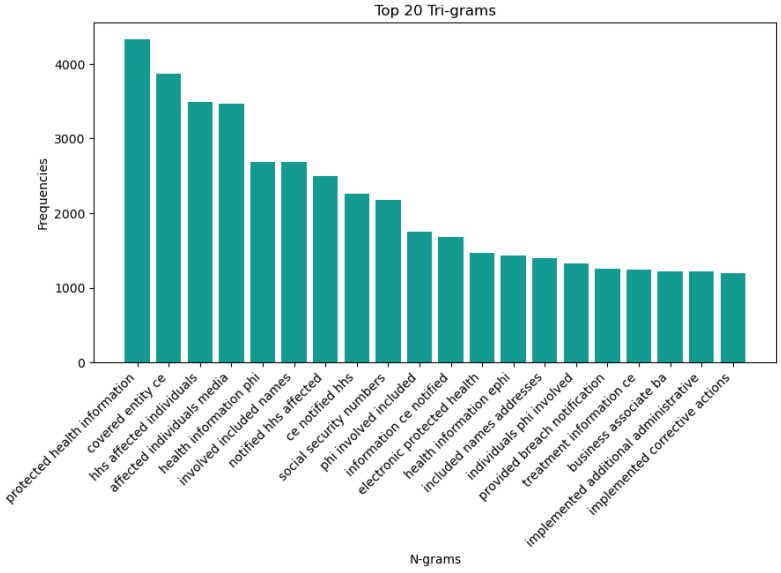
**Descriptive Analysis**

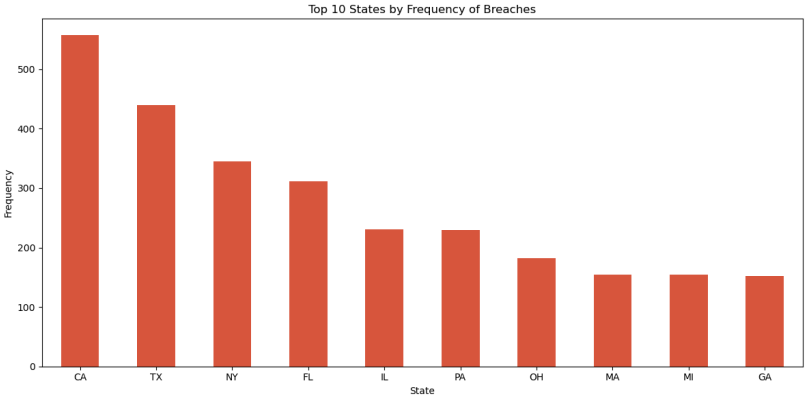


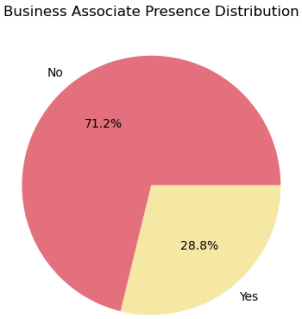

Top 20 Bi-grams
Frequencies
N-grams
affected individuals
health information
protected health
covered entity
entity ce
hhs affected
individuals media
included names
involved included
information phi
dates birth
notified hhs
ce notified
ce implemented
social security
security numbers
phi involved
breach ce
information ce
implemented additional


Covered Entity Type Distribution
Healthcare Provider
72.0%
0.2%
Healthcare Clearing House
13.1%
Health Plan
14.7%
Business Associate


Frequency Distribution of Type of Breach
Frequency
Type of Breach
Hacking/IT Incident
Unauthorized Access/Disclosure
Theft
Loss
Improper Disposal
Other


Top 20 Tri-grams
Frequencies
N-grams
protected health information
covered entity ce
hhs affected individuals
affected individuals media
health information phi
involved included names
notified hhs affected
ce notified hhs
social security numbers
phi involved included
information ce notified
electronic protected health
health information ephi
included names addresses
individuals phi involved
provided breach notification
treatment information ce
business associate ba
implemented additional administrative
implemented corrective actions


Top 10 States by Frequency of Breaches
Frequency
State
CA
TX
NY
FL
IL
PA
OH
MA
MI
GA


Business Associate Presence Distribution
No
71.2%
28.8%
Yes