CalPolyPomona

PROACTIVE DATA BREACH RISK MANAGEMENT IN HEALTHCARE

FINAL PROJECT

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

Data breaches in healthcare threaten patient privacy and system integrity. With the rise of digital records, the risk of exposing sensitive information like medical histories and insurance details increases. These breaches can lead to identity theft, financial loss, and diminished trust in healthcare providers. According to the Ponemon Institute's "Cost of a Data Breach Report 2023," healthcare has the highest average breach cost at $10.93 million and takes the longest to identify and contain, averaging 329 days. This prolonged exposure heightens the potential harm to individuals and organizations.

Problem Statement

Traditional security measures may overlook subtle signs of emerging threats due to the complexity and volume of data. This research aims to provide proactive breach risk management by leveraging text analytics, offering a powerful approach to analyze the data breaches pattern and characteristics from unstructured textual data. Text mining techniques such as logistic regression, topic modeling using advanced NLP methods like BERT, LDA, NER can uncover hidden patterns and anomalies, helping understand breaches in protecting sensitive data.

Research Questions

1. What patterns and characteristics are most common in healthcare data breaches based on textual analysis?
2. How can different analytical techniques help in categorizing and understanding the severity and types of healthcare data breaches?

Background Research

Text mining and Natural Language Processing (NLP) significantly enhance healthcare data security. Mehta et al. (2019) demonstrated that combining NLP with machine learning algorithms can detect anomalies in electronic health records (EHRs), improving early breach detection. Zhang et al. (2020) found that specific keywords and patterns often precede breaches, stressing the need for continuous monitoring and real-time analysis.State regulations vary, affecting breach prevention and response. For example, California's CCPA imposes substantial fines and requires robust security and prompt reporting, whereas Alabama has less stringent regulations and lower fines. Federal laws like HIPAA set nationwide standards but allow stricter state measures, causing variability in enforcement.Integrating text mining insights with regulatory understanding underscores the need for a comprehensive approach to healthcare data breach management, combining advanced analytics with strict compliance.

Dataset

The primary dataset for this project is a collection of healthcare breach reports. The dataset includes *5,370 rows* and *9 columns*. The data was collected from publicly available breach reports submitted to the *Department of Health and Human Services (HHS) Office for Civil Rights (OCR)*.

1. Key Variables:

Name of Covered Entity: The organization that experienced the breach, State: The state where the breach occurred, Type of Breach: The nature of the breach (e.g., theft, loss, hacking), Individuals Affected: The number of individuals whose data was compromised, Breach, Submission Date: The date the breach was reported, Web Description: A textual description of the breach incident.

2. Data Collection Method:

The data was collected from the *OCR breach portal*, which provides detailed reports on healthcare data breaches affecting 500 or more individuals. The dataset was downloaded as a

CSV file and includes breaches reported from various healthcare organizations across the United States.

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.

3. Data Preprocessing and Cleaning Stages

Preparing the data is crucial to ensure accuracy and consistency. This involves:

1. **Standardizing Column Names:**
   - Replacing spaces and hyphens with underscores.
   - Converting all letters to lowercase.
2. **Data Type Conversion:**
   - Converting breach_submission_date to datetime for time-based analyses.
   - Ensuring name_of_covered_entity, type_of_breach, state, and web_description are strings.
3. **Handling Missing Values:**
   - Calculating the proportion of missing values (0.00012%).
   - Using mode imputation for categorical columns.
   - Forward-filling missing breach_submission_date values.
   - Replacing missing values in individuals_affected with the column’s average.
   - Removing duplicate rows.
4. **Data Processing:**
   - Removing stop words using NLTK.
   - Removing non-alphanumeric characters and converting text to lowercase.
   - Lemmatizing tokens with WordNetLemmatizer.

    - Tokenizing cleaned text.

5. **Text Data Preparation:**
    - Building a text corpus from the web_description column.
    - Ensuring all elements in web_description are strings for uniform processing.

4. Descriptive Results of Textual Variables

Analyzing the healthcare breach reports revealed key insights:

- **Most Frequent Words:** Common terms like "information," "individuals," "health," "affected," "breach," "security," and "provided" indicate the core focus areas of the reports.
- **Bi-grams and Tri-grams:** Highlight central themes such as data security, types of information affected, and common breach scenarios.

Top Bi-grams

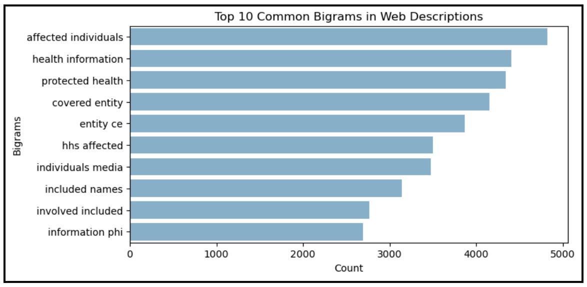


Figure 1. Top 10 Common Bigrams in Web Descriptions

Top Tri-grams

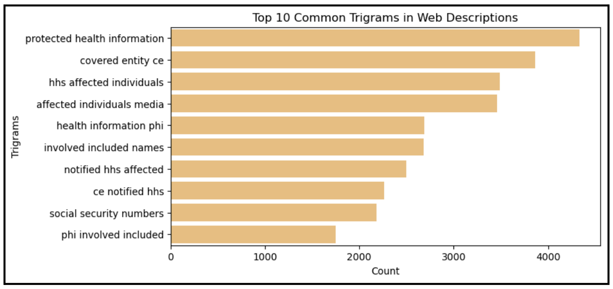


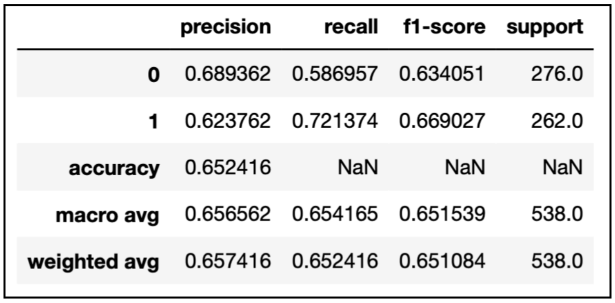
Figure 2. Top 10 Common Trigrams in Web Description

Methodology

1. Logistic Regression

The logistic regression model we developed is designed to predict whether a healthcare data breach is significant or not, based on a 75th percentile threshold for the number of individuals affected.

The threshold used in this model is the 75th percentile of the individuals_affected column. This means that the threshold separates the top 25% of breaches (in terms of the number of individuals affected) from the rest. The 75th percentile is chosen to identify breaches that are in the upper quartile, representing a higher impact in terms of the number of affected individuals. This approach is particularly useful in identifying and prioritizing more severe incidents that might require urgent attention or specific regulatory reporting.



	precision	recall	f1-score	support
0	0.689362	0.586957	0.634051	276.0
1	0.623762	0.721374	0.669027	262.0
accuracy	0.652416	NaN	NaN	NaN
macro avg	0.656562	0.654165	0.651539	538.0
weighted avg	0.657416	0.652416	0.651084	538.0

Table 1. Logistic Regression Model Results

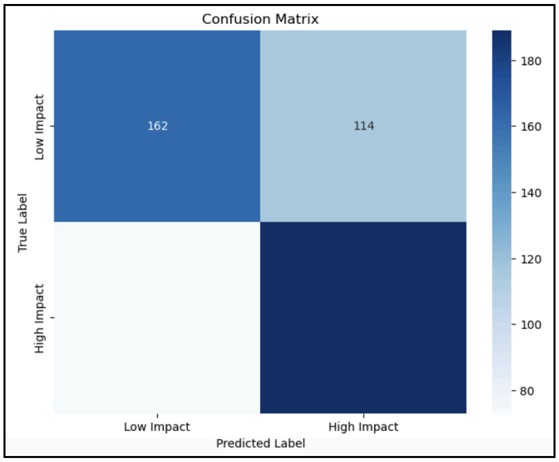


Figure 3. Confusion Matrix

The logistic regression model effectively identifies low-impact breaches, with a precision of 70% and a recall of 59%, minimizing false positives and unnecessary alerts. However, its performance for high-impact breaches is less robust, with a precision of 64% and a recall of 74%. The confusion matrix shows that out of 266 actual high-impact breaches, only 74% were correctly identified, with 26% misclassified as low-impact to minimize damage, maintain data protection compliance, and safeguard patient data integrity.These results are crucial for shaping risk management strategies, highlighting areas where security measures need strengthening. Accurate identification of breach severity is essential for timely responses, and detection shortfalls in high-impact cases emphasize the need for vigilant monitoring and rapid response protocols. Insights from the model can help healthcare organizations refine incident response plans, ensuring high-impact breaches are swiftly addressed

2. Topic Modeling: NER Method

The NER analysis on 930 documents identified 5,735 entities, providing a multi-dimensional view of data breaches, particularly highlighting vulnerabilities in the healthcare sector, regulatory challenges, and geographical concentrations.

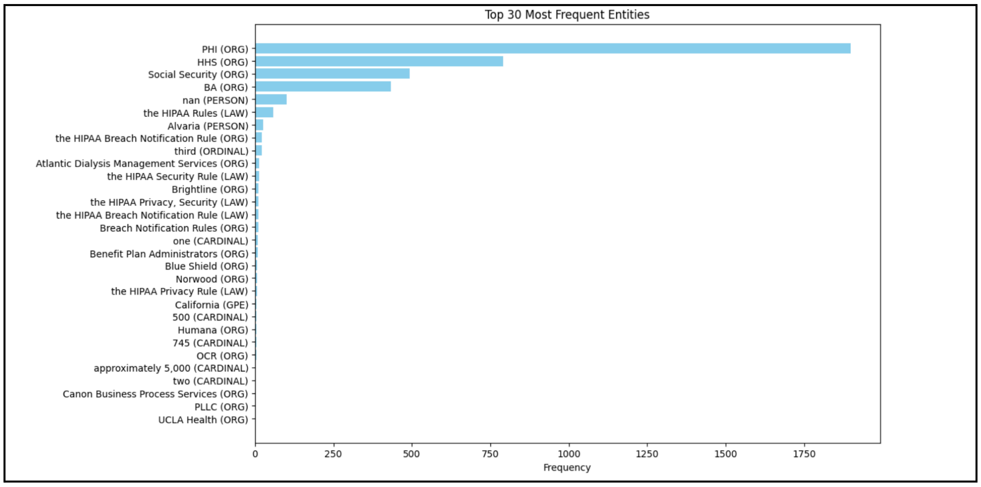


Figure 4. Top 20 Most Frequent Entities

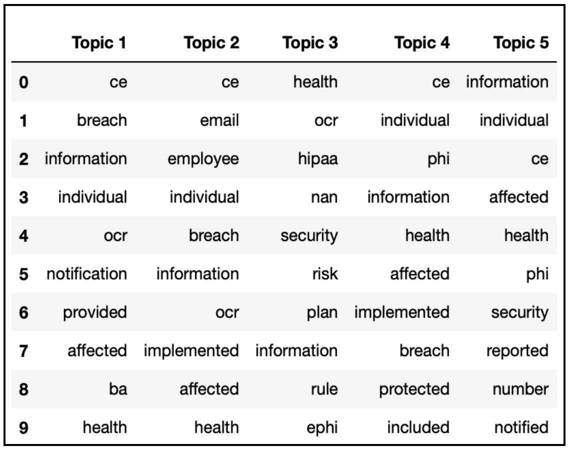
- **Focus on PHI:** Protected Health Information (PHI) is significantly targeted in data breaches, evidenced by its occurrence approximately 1,897 times, underscoring the necessity to secure this sensitive data.
- **Regulatory Compliance Pressure:** The frequent mentions of the Health and Human Services (HHS) and HIPAA-related regulations indicate substantial compliance pressures on organizations, emphasizing the challenge of adhering to strict security mandates.
- **Vulnerability of Healthcare Entities:** Entities like Social Security and Business Associates, as well as major healthcare providers like Atlantic Dialysis Management

Services, Blue Shield, Humana, and UCLA Health, are often mentioned, highlighting their susceptibility to data breaches.

- **Geographical and Quantitative Aspects:** California is noted for a high number of reported breaches, possibly due to its strict data protection laws or large population. References to specific quantities like 500 or approximately 5,000 indicate the scale of breaches and their significant impact.
- **Complex Regulatory Landscape:** The frequent appearance of various HIPAA rules suggests that compliance is a critical component of data risk management, with breaches often linked to the complexities of adhering to these regulations.

3. Topic Modeling: LDA Method

Using the preprocessed text, we constructed a document-term matrix with the CountVectorizer, setting parameters to exclude terms that were too common or rare across the documents. We then applied LDA, specifying five topics *(Table 1)* to uncover underlying patterns in the data. Each topic was defined by a set of top words that frequently co-occur, providing insight into prevalent themes such as compliance with regulations, security measures, and the involvement of the Office for Civil Rights (OCR) in investigations. This methodical approach allowed us to systematically identify and categorize key concerns and issues within the dataset.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
0	ce	ce	health	ce	information
1	breach	email	ocr	individual	individual
2	information	employee	hipaa	phi	ce
3	individual	individual	nan	information	affected
4	ocr	breach	security	health	health
5	notification	information	risk	affected	phi
6	provided	ocr	plan	implemented	security
7	affected	implemented	information	breach	reported
8	ba	affected	rule	protected	number
9	health	health	ephi	included	notified

Table 2. Top Words in LDA-Identified Topics

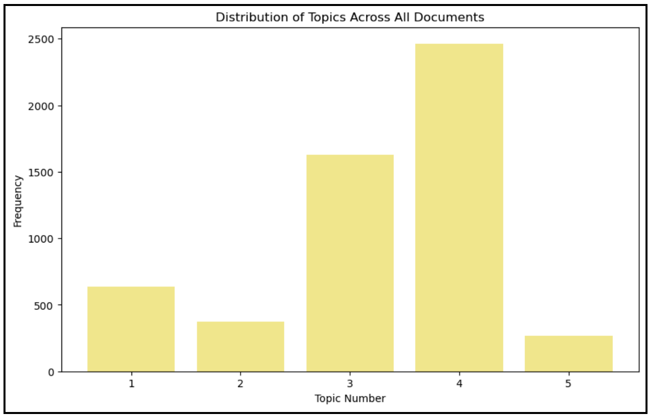
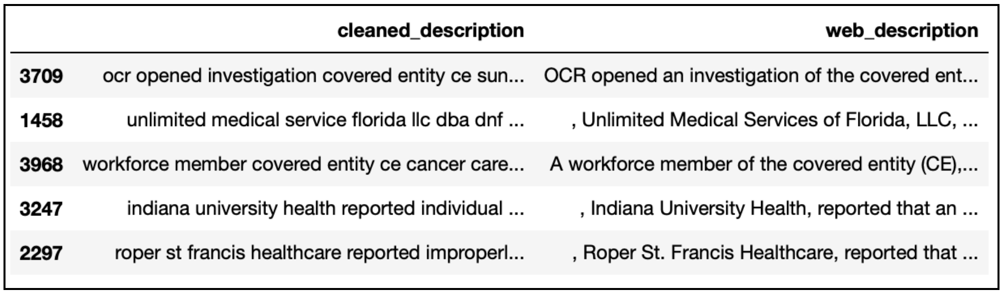


Figure 5. Distribution of Topics Across All Documents

The bar chart shows the distribution of topics across all documents in the dataset. Each bar represents the frequency of a particular topic, with the topic numbers corresponding to the topics identified earlier.

Topic 4 appears to be a significant theme in the dataset, as indicated by its frequency. It revolves around incidents involving covered entities (CEs) where there has been a breach of information related to individuals. The frequent mention of terms like "OCR" (Office for Civil Rights) highlights the *regulatory oversight* and potential investigations or compliance actions that follow such breaches, particularly those involving Protected Health Information (PHI). This topic also underscores the critical importance of proper *notification procedures*, as indicated by words like "notification," "provided," and "affected," which point to the required response actions such as informing affected individuals about the breach. Additionally, the focus on "health" and "PHI" underscores the sensitive nature of the data involved, emphasizing the need for transparency and stringent protective measures.



	cleaned_description	web_description
3709	ocr opened investigation covered entity ce sun...	OCR opened an investigation of the covered ent...
1458	unlimited medical service florida llc dba dnf ...	, Unlimited Medical Services of Florida, LLC, ...
3968	workforce member covered entity ce cancer care...	A workforce member of the covered entity (CE),...
3247	indiana university health reported individual ...	, Indiana University Health, reported that an ...
2297	roper st francis healthcare reported improperl...	, Roper St. Francis Healthcare, reported that ...

Table 3. Top 5 Descriptions Associated with Topic 4

These descriptions reveal the significant role of regulatory bodies in overseeing compliance and ensuring that healthcare entities adhere to legal requirements, particularly in the event of a data breach. It also highlights the necessity for healthcare providers to have robust data protection protocols and clear response strategies to manage breaches effectively and protect sensitive health information.

4. Topic Modeling: BERT Method

The BERTopic modeling, which leverages transformer-based techniques, provided a nuanced understanding of the prevalent themes within the dataset. The model identified 16 distinct topics, each represented by a set of representative words and documents. The analysis revealed that the most frequent topics revolved around the involvement of Business Associates (BAs) in data breaches, the pervasive threat of ransomware attacks, and the critical role of email as a vulnerability. Our group also found out that the intertopic distance map further illustrated the relationships between these topics, highlighting clusters of closely related themes. We applied the BERTopic modeling to provide a comprehensive and insightful overview of the data breach landscape, facilitating a deeper understanding of the challenges and vulnerabilities faced by organizations in safeguarding sensitive information.

In addition, the library `topic_model.topics_per_class` illustrates the distribution of various topics across different classes within the targeted column, it reveals:

- Topic Prevalence: Topic 0 (the_ba_and_its) is the most frequent topic overall, being highly represented in Class 0 and Class 8. This suggests that the theme or concept represented by this topic is a significant factor in these classes.
- Class-Specific Topics: Certain topics are more concentrated in specific classes. For example, Topic 2 (the_ba_and_ransomware) is primarily found in Class 4, indicating a strong association between this topic and the characteristics of that class. Similarly, Topic 1 (the_and_ransomware_ce) is mostly associated with Class 2.
- Topic Diversity: The distribution of topics varies across classes. Some classes, like Class 0, exhibit a wider range of topics, while others, like Class 4, are more dominated by a single topic. This suggests that the diversity of topics could be an important factor in differentiating between classes.
- Potential Relationships: The graph hints at potential relationships between classes and topics. For instance, the predominance of Topic 0 in both Class 0 and Class 8 might suggest a connection between these classes.

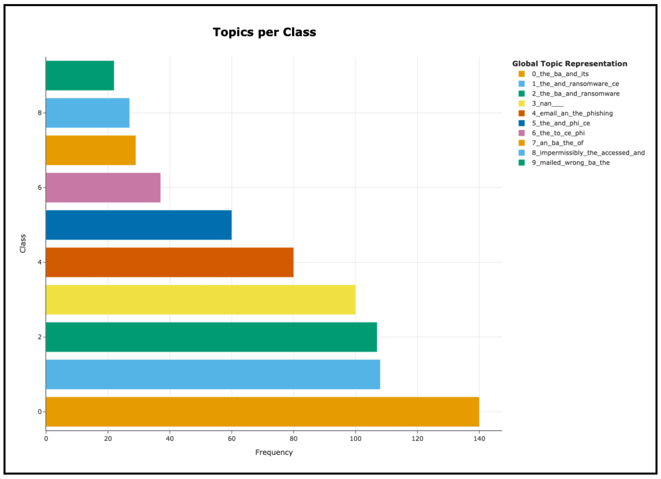


Figure 5. Distribution of Topics Across All Documents

Discussion of Major Findings

Patterns and Characteristics in Healthcare Data Breaches

Our analysis has revealed distinct patterns and characteristics most common in healthcare data breaches, primarily derived from the textual analysis of the web_description column in our dataset. The frequent occurrences of terms such as "information," "health," "affected," and "security" suggest that breaches often involve the exposure of sensitive personal and health information. Additionally, the prominent mention of "hacking" and "IT incidents" identifies these as the most significant types of breaches, affecting substantial numbers of individuals, particularly in states with major tech industries like California, New York, and Texas.

Utilizing Analytical Techniques for Severity Categorization

Using BERTopic, the technique has proved crucial in categorizing and understanding the severity and types of healthcare data breaches. These methods have facilitated a deeper understanding of breach dynamics by identifying distinct clusters of breaches, such as those prominently involving Business Associates (BAs) and the threat of ransomware. The use of advanced NLP and clustering algorithms has allowed us to highlight the critical role of regulatory compliance and the need for robust data protection protocols.

Limitations

Our study, while comprehensive, encounters several limitations:

1. **Data Completeness:** The reliance on publicly available breach data may not fully represent the landscape as not all breaches are reported.
2. **Complexity and Bias in Data:** Due to the web_description column containing many similar contents, such as the way breaches are described and the action methods for

addressing them, there is a high degree of similarity in the data, which impacts the accuracy of the analytical models used. The high uniformity in descriptions can diminish the ability to detect subtle differences between types of breaches, thus affecting the outcomes of classification and deeper analysis.

Directions for Future Studies

To build on the current research, future studies could take several directions:

1. **Expanding the Dataset:** Incorporating a broader array of data sources, including international breach reports, would provide a more comprehensive global view of data breaches.
2. **Refining Analytical Techniques:** Enhancing the precision of existing models, such as the logistic regression used, could improve the reliability of categorizing breach severity.
3. **Long-term Impact Analysis:** A longitudinal study on the repercussions of data breaches on healthcare entities and patients could offer valuable insights into effective preventive.

References

1. Mehta, P., Patel, V., & Dave, M. (2019). Healthcare Data Analysis Using Text Mining. *Journal of Healthcare Informatics Research*, 4(2), 112-124.
2. Zhang, L., Wang, Y., & Li, H. (2020). Predictive Analytics in Healthcare Security: A Text Mining Approach. *International Journal of Medical Informatics*, 135, 104039.
3. Patel, S., Desai, M., & Singh, R. (2021). Comparative Analysis of Breach Prevention Models in Healthcare. *IEEE Journal of Biomedical and Health Informatics*, 25(3), 836-846.
4. Ponemon Institute. (2023). Cost of a Data Breach Report 2023. Retrieved from Ponemon Institute Report.
5. Health Insurance Portability and Accountability Act (HIPAA) of 1996. Public Law 104-191. Retrieved from HIPAA 1996.
6. California Consumer Privacy Act (CCPA), 2018. Retrieved from CCPA 2018.
7. Alabama Data Breach Notification Act, 2018. Retrieved from Alabama Data Breach Notification Act 2018.
8. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084. Retrieved from Sentence-BERT.
9. Grootendorst, M. (2020). BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. arXiv preprint arXiv:2010.00625. Retrieved from BERTopic.
10. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297). Retrieved from KMeans Clustering.

11. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. Retrieved from PCA Review.