

Article

Exploring Key Issues in Cybersecurity Data Breaches: Analyzing Data Breach Litigation with ML-Based Text Analytics

Dominik Molitor¹, Wullianallur Raghupathi^{1,*} , Aditya Saharia¹ and Viju Raghupathi² 

¹ Gabelli School of Business, Fordham University, New York, NY 10023, USA; dmolitor@fordham.edu (D.M.); saharia@fordham.edu (A.S.)

² Koppelman School of Business, Brooklyn College, City University of New York, Brooklyn, NY 11210, USA; vraghupathi@brooklyn.cuny.edu

* Correspondence: raghupathi@fordham.edu

Abstract: While data breaches are a frequent and universal phenomenon, the characteristics and dimensions of data breaches are unexplored. In this novel exploratory research, we apply machine learning (ML) and text analytics to a comprehensive collection of data breach litigation cases to extract insights from the narratives contained within these cases. Our analysis shows stakeholders (e.g., litigants) are concerned about major topics related to identity theft, hacker, negligence, FCRA (Fair Credit Reporting Act), cybersecurity, insurance, phone device, TCPA (Telephone Consumer Protection Act), credit card, merchant, privacy, and others. The topics fall into four major clusters: “phone scams”, “cybersecurity”, “identity theft”, and “business data breach”. By utilizing ML, text analytics, and descriptive data visualizations, our study serves as a foundational piece for comprehensively analyzing large textual datasets. The findings hold significant implications for both researchers and practitioners in cybersecurity, especially those grappling with the challenges of data breaches.

Keywords: cybersecurity; data breach; machine learning; text analytics; litigation case



Citation: Molitor, D.; Raghupathi, W.; Saharia, A.; Raghupathi, V. Exploring Key Issues in Cybersecurity Data Breaches: Analyzing Data Breach Litigation with ML-Based Text Analytics. *Information* **2023**, *14*, 600. <https://doi.org/10.3390/info14110600>

Academic Editors: Eftim Zdravevski, Petre Lameski and Ivan Miguel Pires

Received: 29 September 2023

Revised: 30 October 2023

Accepted: 2 November 2023

Published: 5 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Data breaches have become increasingly common as businesses become more reliant on the Internet and digitized processes [1–3]. Simultaneously, the costs associated with a data breach have also increased. In 2006, the average expenditure on addressing a data breach in the U.S. stood at approximately USD 3.5 million, which soared to about USD 8.64 million by 2020, marking an increase of over 140% within a span of 14 years [4]. As data breaches can lead to public relations disasters and even result in the termination of executive tenures, their prevention, timely detection, and adept management have become a high priority for managers. In a CEO survey conducted in late 2019, half of the participating American CEOs were “extremely concerned” about cyber threats to their business and listed such threats as the single biggest danger to their companies [5]. Among cyber threats, data breaches are rated as the most important issue for security managers, as a recent Delphi study of CISO priorities has shown [3,6].

Recognizing the importance of understanding and preventing data breaches, this exploratory study adopts a novel approach, namely, applying machine learning and text analytics to the interpretation of data breach court cases. The purpose is to extract insights into the nature and dimensions of data breaches from this data source, as there is a recognized paucity of data about data breaches.

Broadly defined, a data breach constitutes the deliberate or accidental exposure of guarded or personal information to a suspicious domain [1,7–9]. Alternatively, a data breach is the conscious or unintended revelation of private information inappropriately to unlawful parties [1]. These occurrences could have severe consequences for a company that could result in very large costs. The incidents can have various causes, and even

though data breaches are hard to prevent, they are not so difficult to anticipate [1,2,10]. They usually result from improper encryption and stolen credentials, one of the simplest and most common ways for cyber attackers to hack or infiltrate a system when it is made of predictable and easy-to-decrypt passwords [11,12]. They also come from configuration errors when a software's technical vulnerabilities are exploited by attackers [1]. This usually happens after a software provider discovers a vulnerability, urging their clients to apply a fix. If not applied promptly, attackers may take this opportunity to exploit the weakness to steal customer data [11,12]. Furthermore, malware is also a way cyber attackers can access confidential data. All it takes is to install malware in a piece of software that contains a known vulnerability and exploit the rewards [1,13]. Additionally, these data breaches can also come from inside the company. On the intentional side, employees may be tempted by the financial gain of selling data on the dark web or feel resentment towards the company, accessing the organization's systems for nefarious purposes. On the unintentional side, employees may just commit a mistake that results in a data breach, such as including the wrong person when sending an email, attaching the wrong document, or losing a laptop [1,14].

A study conducted by IBM in 2020 on the costs of data breaches suggested that the average overall cost of a data breach had hit approximately USD 4 million [4]. In the past decade, several well-publicized data breaches have brought increased awareness to the consequences of such breaches. Chief among them is the breach of Target Corporation's network in 2013 resulting in the theft of approximately 40 million credit card data and 70 million customers' individually identifiable data. Target estimated the losses to be approximately USD 248 million [15]. Likewise, Yahoo reported in 2016 that at a minimum, at least 500 million accounts had been pilfered in 2014 in a data breach presumably sponsored by a country [16]. In yet another study conducted by Intel Security, it was reported that in-house employees were responsible for 43% of corporate data breaches. However, nearly half of these breaches were inadvertent [17]. The provocations for internal breaches are numerous, key among them being corporate espionage, grudges with employers, or financial motivation.

On the other hand, internal accidental breaches occur primarily from inadvertent causes such as lack of security, governance, non-compliance with policies and procedures, or lapses in oversight [1]. Overall though, a majority of breaches are not reported or disclosed [18–20]. Data collected by the Privacy Rights Clearinghouse (PRC) in 2018 shows 8137 publicly disclosed data breach incidents since 2005 with a total of approximately 10.4 billion breaches. Breaches were mostly prevalent in various businesses with approximately 2397 incidents and a total of 9.8 billion breaches [21]. Interestingly, breaches in businesses dominate over breaches in other sectors (95%). In addition to actual monetary losses, considerable intangible and indirect costs were likely to occur. These may include the cost of damage to the company image, loss of the confidence of the customer, litigation costs, and others [22]. For example, the recent data breach at Equifax cost the company hundreds of millions of dollars and created problems for consumers and other stakeholders [23]. Therefore, it is imperative for the viability and profitability of enterprises to proactively prevent or mitigate such data breaches [12].

As data volume grows exponentially in the era of "big data" opening opportunities of increased number of data breaches, developing, and implementing prevention and mitigation strategies are of paramount importance to businesses. However, one needs to identify and understand the various types of data breaches, which in turn requires more granular research studies investigating both quantitative and qualitative variables [19].

Our motivation for this study is derived from multiple angles: First, data breaches are occurring at an increasing rate in all walks of life [21] and have deeply penetrated all aspects of businesses, while simultaneously, the various stakeholders, (e.g., customers, executives, patients, regulators, etc.), want to be proactively engaged in prevention and mitigation and to be informed; second, what follows is that various stakeholders including organizations, governments, individuals are attempting to respond aggressively and proactively to the

threat of data breaches; third, businesses are proactively using various social media, such as X and Facebook, to communicate with the public regarding data breaches; and fourth, it is generally acknowledged that research in cybersecurity in general and data breaches, in particular, is still scant and incomplete.

Eliciting the nature and dimensions of data breaches from the perceived authentic source of data breach litigation cases in the courts would provide insight into the data breach phenomenon. Our analytic research applies machine learning-based text analysis to extract from and scrutinize the more significant data breach particulars disclosed in the court case filings. While prior research has, in general, explored publicly available news reports, press releases, and voluntarily disclosed databases, we examine data breach details contained in the various court cases and filings at various levels and jurisdictions in the United States. It is important to note that while court documents are official and authentic, they are narrative, textual documents, requiring advanced computing methods such as machine learning to decode them. Therefore, there is an element of subjectivity in the analytical process. Analyzing and gaining insight into data breaches by pulling key insights and issues and from this legitimate source would reveal the nuances of data breaches. In analyzing data breach-related legal cases, we focused primarily on the categories and specific types of data breaches, the relevant statutes, etc.

We supplement existing research on cybersecurity and data breaches in several ways. First, our research examines the case documents (e.g., affidavits) submitted to the courts. Therefore, the sources of our insights are more convincing. Second, we utilize the more contemporary data found in the legal documents, the latest year being 2021, and include an entire decade worth of data that are quite substantial. By conducting an extended longitudinal study and including a larger sample of data breach cases, we broaden the focus of the research to conduct analysis at numerous levels. Further, this permitted us to emphasize a longitudinal understanding of the trends in key topics. In addition, the data set made possible the exploration of the multiple aspects that constitute a data breach, namely, the types of breaches, stakeholder characteristics, the application of statutes, monetary issues, and others. Third, we contribute modestly to our understanding of how NLP, machine learning, and text analytics can be applied to very large text data sets in the data breach and legal domain. Fourth, this study examines data breaches from a legal perspective, focusing on the parties involved in the litigation. As companies, regulatory bodies, watchdog organizations, nonprofits, and NGOs can utilize the insights gleaned from this research, it may be anticipated to foster enhanced detection, mitigation, and prevention strategies against data breaches, consequently bolstering the overall cybersecurity posture. In this context, we critically analyze the court case filings pertaining to data breaches to unravel the core concerns and issues that arise, especially in relation to legal frameworks and the perspectives of the litigants.

The remainder of this paper is organized as follows: Section 2 describes data breaches, data breach litigation, and ML-based text analytics. Section 3 discusses the methods that were applied, the corresponding results, and the analysis; and Section 4 offers a brief discussion. Scope and limitations are highlighted in Section 5 while conclusions and future research directions are offered in Section 6.

2. Background

2.1. Data Breaches—Industry Reports and Empirical Studies

Current research into data breaches is primarily published in white papers and reports by a range of stakeholders, including consulting firms, vendors, and regulatory agencies [24]. For example, the “Verizon Breach Investigation Report” has been published every year since 2000. The report sheds light in the aggregate on prior cyber security occurrences [24]. The report details routine attack trends, threat agents, hacker impulses, breach identification techniques and history, and recent hacking/malware shifts. It is reported that the source of information on data breaches originates in cases explored by Verizon or a collaborating organization such as a think tank or consulting company [24]. In addition,

Verizon has traditionally put out its “Data Breach Digest”. The Global Security Report, an annual report, is produced by Trustwave. This report showcases the premier cyber security threats and attack directions [25].

In addition to the reports, a few case studies have been published regarding data breaches. For example, Manworren et al. [26] discuss the Target data breach case. In another type of study, Rashid et al. [27] developed a framework that is inclusive of the different steps in a data breach event. They further discuss breach identification and resolution strategies. In their research, Collins et al. [28] conducted a comprehensive literature review to evaluate the contemporary status of data breaches in various organizations. The authors further analyzed the reported data breaches over a six-year period by Privacy Rights Clearinghouse. The sample consisted of approximately 2219 data breaches revealed between 2005 and 2010 [28].

The present-day research into data breaches is generally reported in white papers and reports put out by various consulting companies, vendors, and regulatory agencies [24]. The study analyzed four key metrics: breach type, reporting organization, year of disclosure, and the location of the breach [28]. In another study, Posey Garrison and Ncube [29] analyzed data regarding data breaches over a five-year period to investigate the potential association between the type of data breach and the organization. They categorized the data by breach and institution type, record size, and state. The study found that typical breach types led to either stealing or exposure of data [29]. The authors concluded that educational organizations had more breaches, typically by hacking or via exposure. Also, the number of insider breaches was less than that of the other breach types [29].

Ayyagari [30] conducted content analysis of approximately 2633 different data breaches that collectively resulted in a loss of more than 500 million individual records. The period of study was 2005–2011. The study points out that data breaches were a major challenge for businesses. However, the study found that there was a shift from hacking to the human dimension regarding data breaches. An unsettling finding was that security policy (or lack thereof) was a key element in data breach occurrence [30]. In their research, Khey and Sainato [31] utilized a six-year sample of data gathered by Privacy Rights Clearinghouse, to study the potential association between data loss and the location of the breach.

Zadeh [32] attempted to quantify and categorize the severity of a data breach using several metrics such as data asset type, account data, or financial data about the breach. The information on breaches retrieved from S&P 500 corporations was further studied to derive mitigation strategies. Hacking and malware were the most prominent data breaches [32].

In their research, Hammouchi et al. [33] examined publicly available information on over 9000 data breaches in the 2005–2018 period. It is believed during this time, approximately 11.5 billion individual records were stolen resulting in huge financial and technical impacts. The authors also provided insight into the breach type, the most targeted organizations, and the evolution of hacking over time. This study concluded that while hacking was on the rise, breaches caused by human errors declined as organizations implemented security policies and procedures [33]. Shu et al. [34] investigated the data breach event at Target from a legal angle. Specifically, Smith [35] examined data breaches in healthcare organizations. The goal of this study was to determine the correlation between several metrics including breach type, location, organization type, and the number of individuals impacted. The study found that 70% of the intrusions were in the healthcare providers’ sites and they mostly had to do with digital systems and data.

Holtfreter and Harrington [36] used a novel model to gain insight into the different data breach types of approximately 2280 breaches and over 512 million corresponding jeopardized records within the U.S. They too drew their data from the Privacy Rights Clearinghouse for the period 2005–2010. Their research revealed that while the annual number of data breaches and the corresponding number of compromised records generally increased over the study period, these trends fluctuated inconsistently from one year to the next. Neto et al. [37] created a database of available at <https://www.databreachdb.com> (accessed on 1 October 2023). Insights gleaned from this database showed that the number

of data records breached had increased from approximately four billion in 2018 to more than 22 billion in 2019. According to their paper, this increase transpired even though robust initiatives from government regulatory entities have attempted to formulate and execute tighter policies and procedures. Examples include the General Data Protection Regulation (GDPR) that became effective in Europe in the month of May in 2018. However, the proactive enforcement of regulation in Europe has resulted in the disclosure of the prevalence of a larger number of data breaches there compared to the disclosures in the U.S. (>10,000 data breaches publicly reported since 2018 in the U.S. vs. >160,000 reported in the E.U.).

2.2. Data Breaches—Theoretical, Conceptual, and Policy-Focused Studies

This section explores the theoretical, conceptual, and policy-focused studies that offer nuanced insights into data breaches from various disciplinary lenses. While some studies focus on broad overviews of data security, others concentrate on specific sectors, such as healthcare. For instance, Cheng et al. [1] carried out a comprehensive review of the literature to gain insight into what they termed data leakages. Similarly, McLeod and Dolezel [38] conducted a study of data breaches in healthcare organizations and found that exposure level, the nature of security in place, and other organizational factors may cause data breaches. Algarni and Malaiya [39] studied the cost side of breaches. They examined how data breach costs were assessed and calculated and what factors were associated with these costs. Kafali et al. [40] examined data breaches from a policy perspective. They investigated the association between specific policies and the related data and assessed the gap between the two. Surprisingly, they found that the number of unintentional revelations of data were about the same as intentional ones. In a comprehensive study, Sen and Borle [41] examined data breaches from multiple perspectives, applying various theories to identify and assess the risks associated with these incidents. Interestingly, their research indicated a positive correlation between investment in information technology (IT) and the risk of data breaches. They concluded that greater utilization of technology increases the likelihood of experiencing a breach.

Hall and Wright [42] analyzed the breaches between 2014 and 2018 and their overall conclusion was that cyber-attacks leading to breaches occur in nearly every industry, more so in the healthcare industry. Schlackl et al. [3] reviewed 43 articles on data breaches prior to their occurrence and 83 for post-occurrence impact. Their theoretical framework encompassed a variety of lenses, ranging from viewing a data breach as an organizational crisis to exploring theories related to criminality and privacy. Romanosky [43] put together an unusual dataset of over 12,000 breach incidents documented between 2004 and 2015 and carried out a comprehensive analysis of the breaches by examining the components and costs by industry type. Despite the ample coverage of data breaches in both the academic literature and the media, there is a surprisingly limited body of research dedicated to thoroughly examining these incidents to accurately assess associated risks and trends.

Against this backdrop, we identify three relevant gaps in the literature. First, although studying data breaches is critical [3,39], empirical work at the granular level is sparse, and we found that most, if not all, studies exclusively reported case studies of major data breaches, or data taken together [43,44]. Second, very few studies have utilized machine learning and text analytics to explore the data. Third, the few data-driven studies that exist are outdated. Our study attempts to fill these gaps.

2.3. Data Breach Litigation

While data breach lawsuits have been around for a while, particularly class action litigation, the number of cases is rising quickly, implying an increase in the number of mega data breaches affecting millions of records and people [45]. For example, there has been a gradual increase in the number of litigation cases in data breaches between 2019 and 2021, with organizations that were frequent targets of cyberattacks being those of the government, healthcare, retail, and technology [46]. Lawsuits are being filed individually and collectively

to obtain legal remedies for injury caused by the intentional theft or unintentional loss of personal data [13]. For example, HCA Healthcare's data breach last summer, revealed in a July 10th press release, unleashed an avalanche of litigation against HCA [47]. At a minimum, twenty-three lawsuits have been filed with respect to this data breach since the press release. According to the filing, HCA patients who were affected by the breach had experienced harm due to the invasion of their privacy and jeopardizing their data information among other grounds for the filing.

Businesses also face the risk of tarnishing their images when the personally identifiable information ("PII") of their clients or customers is lost or stolen during a data breach [48]. Simultaneously, escalating costs of mitigating data breaches have also resulted in huge settlement expenses associated with class action litigation. Hacking and malware have resulted in costs exceeding USD 4.4 billion [49]. For example, a court-approved settlement of the notorious Equifax data breach included approximately USD 380 million for a tentative settlement fund and a further sum of USD 125 million for ad hoc expenses. This breach had resulted in the unauthorized access to personal data of over 147 million Americans (<https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement> (accessed on 1 October 2023)). Likewise, many prominent companies have encountered class action litigation filed by plaintiffs whose data privacy was compromised, for example, under the Fair Credit Reporting Act (FCRA).

Settlement amounts in data breach class action litigation have topped hundreds of millions of dollars: Home Depot (USD 200 Million); Capital One (USD 190 Million); Uber (USD 148 Million); Morgan Stanley (USD 120 Million); and Yahoo! (USD 85 Million) [50]. Plaintiffs can also sue under state laws, such as The California Consumer Privacy Act (CCPA). A nuance to class action litigation is the imposition of fines, both as a penalty and as a deterrence. For example, Marriott was initially fined USD 124 million which was later reduced. The Chinese firm Didi Global was slapped with a USD 1.19 billion fine for violation of China's data protection law. Amazon was fined USD 877 million for violation of the General Data Protection Regulation (GDPR) in Europe [6,49]. Hill and Swinhoe [49] enumerated 36 significant data breach class action lawsuits in 2021. This is significantly higher than the approximately 25 lawsuits filed in 2020 (1/22 <https://www.mofo.com/resources/insights/220104-privacy-litigation-2021-year-review> (accessed on 1 October 2023)). To note, 16 settlements were agreed to in key federal data breach filings in 2021 (<https://www.mofo.com/resources/insights/220104-privacy-litigation-2021-year-review> (accessed on 1 October 2023)).

Our research attempts to shed light on the nature and dimensions of data breaches by analyzing them via the lens of data breach litigation. We used contemporary methods of machine learning-based text analytics of the cases' documents for this purpose. To the best of our knowledge, no empirical research involving lawsuits in data breaches utilizing machine learning and text analytics has been carried out. The goal of this study is to fill this research and policy gap by critically analyzing a representative sample of federal lawsuits associated with data breaches. Westlaw was searched using keywords and the court decisions were downloaded in the form of portable document format (pdf).

2.4. Text Analytics and Machine Learning

Text analytics, an important application of machine learning, has gained attention due to the abundant availability of text-based documents in a variety of domains. Whether it is a corpus of tweets or corporate customer support and feedback narratives, text analytics can analyze these large corpuses to gain insight [51]. Machine learning (ML), natural language processing (NLP), and text analytics can auto-generate clusters, classify, and glean patterns from the text. By applying NLP, text analytics can convert free-form unstructured text into a structured format that is suitable for the application of ML algorithms. Utilizing text analytics, data scientists can identify and evaluate various features that describe the unstructured text. A majority of studies utilizing text analytics studies apply thesauruses consisting of words or phrases with similar meanings [52]. To interpret a large volume of

text, the techniques identify the relative frequency of occurrence of key terms and groupings and assess the comparative significance of key ideas in the text. The primary benefit of text analytics lies in its ability to process large volumes of text data [53].

In the context of this study, analyzing court cases in data breach litigation offers a relevant subject. For one, it includes a large corpus of unstructured text. Additionally, the corpus includes important facts, laws, and precedents of prior data breach cases and presents an appropriate repository to explore. Data breach litigation cases offer a novel source of information regarding data breaches including data breach types, parties (stakeholders) involved in the litigation (e.g., corporations), the cost, the laws that were involved, and prior cases that impact the current case. Therefore, data breach litigation presents an important source of data-on-data breaches worth exploring. The methodology of text analytics has been used in numerous studies on analyzing unstructured textual information, for example, vaccination sentiment tweets, tweets on vaccination [54], legal patent validity [55], health blogs [56], or shareholder resolutions [57]. Considering the benefits of ML, NLP, and text analytics, we used these methods to analyze a large sample of data breach litigation cases.

3. Methods

This exploratory study, therefore, examined data breach litigation cases to elicit the broad data breach types and laws, litigants, and trends by applying machine ML-based text analytic methods. The source of the data breach litigation cases is Westlaw, the legal database (<https://legal.thomsonreuters.com/en/westlaw> (accessed on 1 October 2023)). Westlaw is an online legal research service and proprietary database including thousands of cases, courts, state, and federal laws, and others. The ML-based text analytics approach enabled the distilling and synthesis of a large volume of textual data in a productive manner [58,59]. The insight gained from the study informs corporate executive suite, other relevant parties, cyber security experts and policymakers, consumers, the government, NGOs, and others as to the prevention, mitigation, and implications of current and future data breach occurrences.

The relevant cases used in the study were retrieved from the Westlaw database using the online query search function for the period 1 June 2019 to 31 May 2021 using a combination of keywords such as “cyber security”, “data breach”, etc. The Python selenium package (pypi.org/project/selenium/ (accessed on 1 October 2023)) was additionally used for crawling and preprocessing the pdf files, one file per case. Several cases were duplicated in the results of the query search. Additionally, many of the PDFs were not complete case descriptions with decisions, rather, they were ad hoc filings. Furthermore, many others were irrelevant, for example, “breach” was used in other contexts, such as in river dam breaches. Since the data are in an html structure, it is necessary to eliminate the structural data. For example, there were footnotes repeatedly appearing on each page, such as “Thomas Roberts”, etc. We extended the stop words processing to include these frequently appearing footnotes. Simultaneously, legal terms such as “court”, “defendant”, and “attorneys” were also included in the stop words. Furthermore, we removed unrelated hyperlinks, notations, and symbols to ensure the relevance and quality of the documents’ content.

A total of 698 legal cases were available for analysis after eliminating irrelevant cases, missing/incomplete content, or non-functional links. These cases were primarily from the District Courts, Courts of Appeals, and the Supreme Court. The 698 cases were then converted into text strings. Figure 1 outlines our methodology and machine learning-based text analytic methods used in this study.

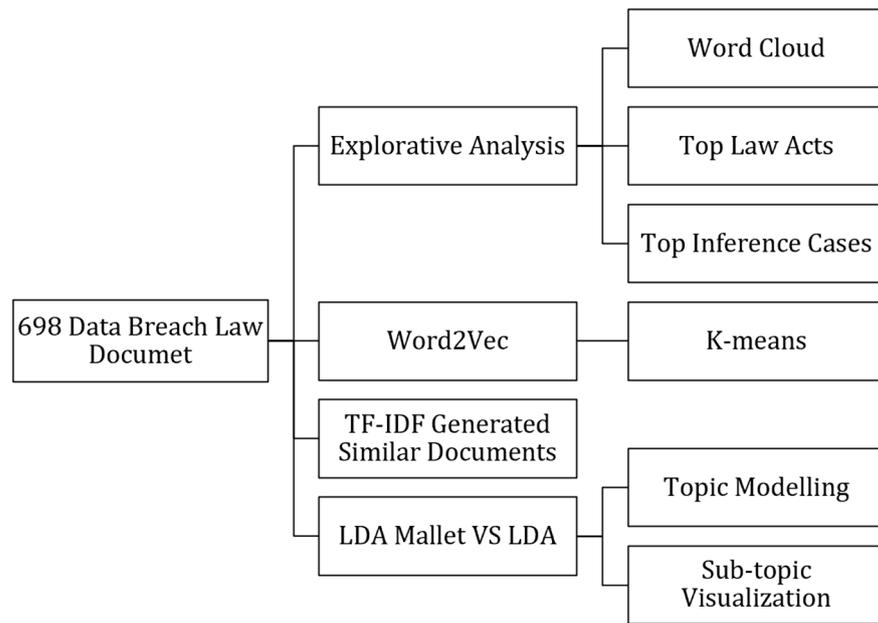


Figure 1. Methodology.

Figure 2, a bar chart shows the most frequent words that appeared in the cases’ narratives. The words “data” and “breach” appeared 10,387 and 10,150 times, followed by the word “work” appearing 8877 times. “Injury” shows up 8052 times and “violate” 6373 times. Next, “credit” and “bank” are also popular words in data breach cases, appearing 5621 and 3168 times. Lastly, “system” is also a frequent word that appears 2934 times. Together, these words indicate that many of the data breach litigation cases dealt mostly with the banking system.



Figure 2. Keyword frequency in data breach.

Text Analytics

What follows is the utilization of ML-based text analytics to explore data breach cases in-depth. The corpus of text data was pre-processed before the application of various algorithms and methods. The content of each case in the sample was saved as a text file. Following this, vectorization was performed. Some of the cases were deleted for being vague or lacking relevance to data breach, and/or unworking links. Predictive modeling conducted on text data faces multiple obstacles in the analytical process. First, textual data lacks the ability to serve as input to numerous quantitative models. To overcome this, an

NLP system was applied to convert the text content into discreet elements as the next step in the methodology. Second, text-based corpuses are larger in size compared to data sets that are of numeric type. A robust and viable model, therefore, warrants the retrieval of the most promising and relevant pieces of data for the problem-solving process.

The initially scraped and raw summary portions were converted into ordinary text via the deletion of numbers, punctuations, spaces, and standard stop words using the Natural Language Toolkit (NLTK) in the data pre-processing step. Following this, the cleaned text was transformed lower case utilizing the NLTK and TextBlob (<https://textblob.readthedocs.io/en/dev/> (accessed on 1 October 2023)).

Next, the lemmatization method was implemented to metamorphose the words into their granular form (for example, “breaching” and “breached” were substituted with “breach”). Lemmatization groups together the assorted similar variations of a word, thereby enabling them to be studied as a unary term and bringing contextual meaning to the terms. The maximum number of features allowed was restricted to around 4000. Tokenized words included those with more than four characters. Latent Dirichlet Allocation (LDA) was then applied to identify primary themes and their dispersion in the volume of data [58,60]. The LDA allows exploration of large corpus of text without the need for an a priori list of terms. In other words, it facilitates the garnering of a collection of concepts without pre-defined notions. Applying this in the context of data breaches, the method allows for a more comprehensive overview into the content of data breach litigation cases, when compared to other methods.

To improve the accuracy of our topic modeling, we also used the LDA Mallet model. Mallet (Machine Learning for Language Toolkit) is an extension of standard LDA and is known for providing more accurate and coherent topics. This model is specifically designed to optimize the coherence score, a metric that quantifies the human interpretability of the identified topics.

The TF-IDF method is also modeled next. The “term frequency-inverse document frequency” (TF-IDF) method was used to calculate the weight of each term to denote its relative significance in the text file(s). In TF-IDF, an information retrieval technique, a weight is allocated to a term’s frequency (TF) and its inverse document frequency (IDF). Each term is given these two scores. The weight of the term is then calculated as a product of these scores. Furthermore, the K-Means clustering model was applied to surface the key data breach concepts. The K-Means clustering is one of the more widely used ML algorithms positioned in the classification of cases based on similarity measures (that is, the distance between cases). It has been traditionally applied in pattern recognition and classification problems. Four clusters were identified using the word cloud package. This was followed by classification with the K-Nearest Neighbor (KNN) algorithm, and the data were split into two portions to measure the classifier success (training and testing).

4. Results and Analysis

This section presents the key results of the ML-based text analytics conducted on the data breach cases. First, word cloud analysis was conducted. The top keywords and word clouds were generated to obtain a high-level overview of the text data using the Text Rank and Term Frequency methods. The word clouds showcase the words most occurring in the sample of cases. These were output by wordclouds.com (accessed on 1 October 2023). The larger the size of the word, the higher its frequency of occurrence in the corpus. This confirms what we knew anecdotally and via exploratory manual reading and analysis of the key implicit issues in randomly picked litigation cases. Subsequently, clustering was undertaken to elicit the primary themes and associated topics to extract core concepts and significant factors within each cluster.

Word2vec was developed to serve as the input to the K-Means algorithm. The two primary machine learning models that were implemented included, specifically, the LDA model and the clustering model. Word2vec models have been developed to feed into the K-Means model, and results have been analyzed by identification of top words and

visualized in a word cloud. To achieve robust results, LDA and LDA Mallet models were applied to explore topics discussed in documents. In addition, we also applied TF–DF to retrieve similar documents.

4.1. Word Cloud

The word cloud maps in Figures 3–5 display the words that are most frequently prevalent in the corpus. Text Rank and Term Frequency models were utilized to develop word cloud maps. Text Rank is a general-purpose, graph-based ranking algorithm utilized in NLP. This algorithm decides the importance of a piece of text within a document, based on the information recursively drawn from the entire document [54,55]. Term Frequency refers to the frequency of a piece of text that has appeared in the whole text. We used three word count methods to identify the most relevant pieces of information in all cases’ files, namely, text rank splitting by page, text rank splitting by case, and term frequency. Since there are several parts in one case, such as background information (holdings, attorneys, and law firms, etc.), the real content (petitioners’ allegations, standing, etc.), and the judges’ opinion, separating cases by page may help eliminate the background, which contains less useful information from a word count perspective. Thus, if the cases are separated by page, one is more likely to retrieve useful information regarding relevant concepts in the cases.

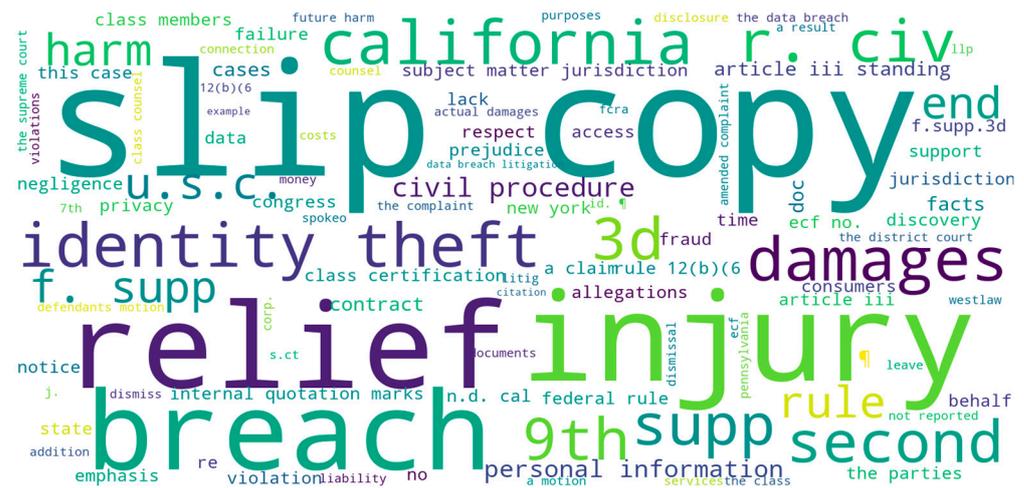


Figure 3. Word cloud of text rank approach split by page.

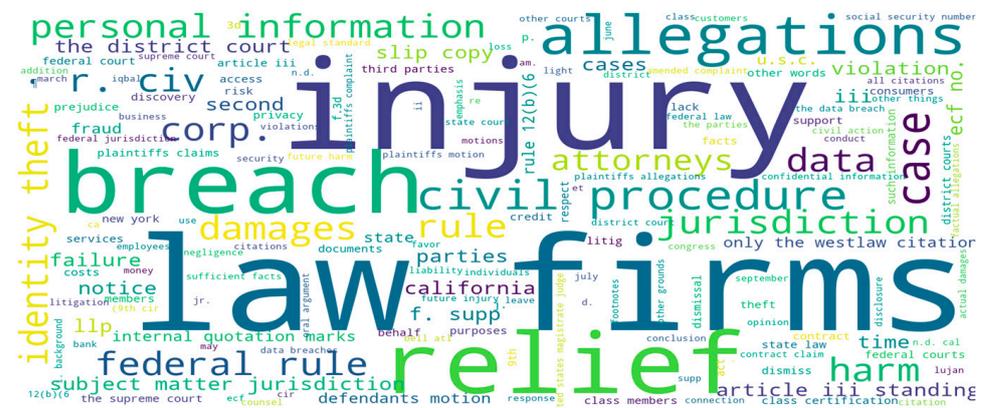


Figure 4. Word cloud of text rank approach split by case.

However, terms generally by themselves do not add to a deeper understanding of data breaches. Therefore, we next conduct topic modeling using LDA and LDA Mallet [58,61,62].

4.2. Topic Modeling

This section describes the application of topic modeling in the identification of the topics that are most prominent in the content of the 698 data-breach-related cases. Topic modeling is the process of applying unsupervised learning to automatically uncover the topics in a set of text documents. The topic structure—the topics, per-document topic distributions, and the per-document per-word topic assignments are hidden structures that need to be covered [58,59,63]. The topic modeling technique usually characterizes the documents as vectors. In the simplest form, such vectors are inclusive of the number of occurrences of each term in the document. However, each vector may contain a greater number of dimensions (each one reflecting one term). It follows logically that we need to reduce the number of dimensions in the resulting vector [64] to cope with the large amount of data. The LDA is therefore applied, and in the resulting vector of LDA, each dimension mirrors a single topic or concept [64].

A topic represents the probability distribution over all the terms that co-occur in the underlying documents [64], and a document itself is a probability distribution over all topics in the corpus [60,65]. It is implied when describing a topic, the researchers choose terms with a particular probability from the pool of terms related to that topic [65]. For instance, when describing the topic of data breaches, terms such as hacking, cybersecurity, stealing, etc., have a higher probability of appearance, while terms such as employee benefits, ethics, or profit gains have a lower probability of appearance. Topics are elicited by recognizing the terms that frequently occur together. The implication is that the more frequently terms occur together in the same document, the more likely it is they describe the same topic. Each litigation case may comprise many topics. The probability distribution of a document can show how significant the recognized topics are in a particular case. After this, the documents were tokenized, that is, partitioned into tokens.

Next, we tokenized the documents—split them into tokens that included keywords and special notation such as punctuation. The text is further refined by converting all characters into lowercase and deleting the special characters and numbers. Utilizing WordNetLemmatizer, the words were then lemmatized, and the process removed routine stop words (i.e., universal words such as articles, conjunctions, pronouns, etc.). The stop word list “English” in the Python-based NLTK package was used. Additionally, terms that were present in less than two documents were deleted. A final check resulted in the elimination of capital nouns. The single most important goal of LDA is to search for in each document the confluence of topics, wherein each topic is denoted by a bag of terms [65]. Therefore, the probability distribution of the confluence of topics differs from the bag of terms [65]. The hyper-parameter α describes the shape of the per-document topic distribution, and the hyper-parameter β describes the shape of the per-topic word distribution [66]. The distributions are estimated by the algorithm using Dirichlet priors [65]. The Gensim 4.1.1 (for Python) and Mallet 202108 (for Java) software are examples of robust platforms for LDA [64].

After a comprehensive preprocessing, the model was executed. First, we ran the widely used LDA version, a generative probabilistic algorithm, specifically appropriate masses of discrete data [60]. This version characterizes case document files as bags of words, and in each bag of words, a batch of keywords is refined collectively and designated as a topic. It is possible to have several topics in a case word bag. Also, the number of topics in each case can be optimized. Numerous parameters can be varied to accommodate the model-building process: the initial maximum number of iterations was set to 50, and the hyperparameters α and β which describe the shape of the per-document topic distribution, and per-topic word were assigned. Each training chunk included approximately 1000 data files from the sample of 698 cases (1 case can be partitioned into several data files). In addition, there are 10 epochs running via the corpus in training. The model also assembled

a list of topics, arranged in descending order of the topics for each word, along with their phi values times the feature length (i.e., word count) as `per_word_topics = True`. A first-cut attempt resulted in 20 topics.

The efficiency of the model was assessed by using the cross-validated (C_V) coherence score as the standard. The C_V coherence score measures the frequency of the co-occurrence of words belonging to the same topic in a corpus. It reflects the semantic similarity between words in a topic. A high coherence score indicates that the words are more semantically similar, and the topic makes more sense. The C_V coherence score is based on four metrics: segmentation of the data into word pairs, calculation of word or word pair probabilities, calculation of a confirmation measure that quantifies how strongly a word set supports another word set, and the aggregation of individual confirmation measures into an overall coherence score [67]. Their phi values multiplied by the feature length (i.e., word count) as `per_word_topics = True` are also calculated. The coherence score has no standard value since it depends on the size of the corpus, but the visualization of the LDA model is not significant.

As Figure 6 shows, the distance between the bubbles represents the topic similarity with respect to the spread of words. The surface of the bubbles represents the prevalence of a topic within the corpus. Since the topic modeling expects to separate each node (bubble) for minimal overlap, the bubbles (sub-topics) need to be more dispersed than overlapping, on average, within the four quadrants (PC1, PC2, PC3, PC4).

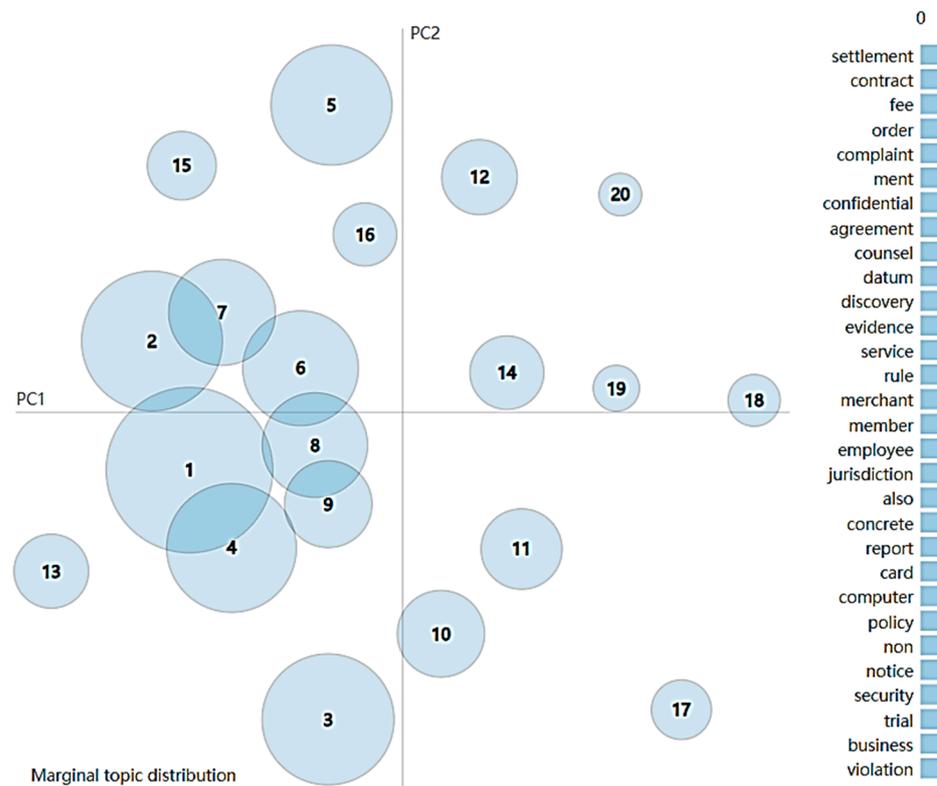


Figure 6. Distribution of LDA model twenty topics.

However, some of the nodes in the cluster tend to group together, diminishing the cluster’s significance. To aim for more precise results, we imported the LDA Mallet model to improve the coherence score. Our aim was to achieve the highest correlation with all existing human topic ranking data, as suggested by Syed and Spruit [67]. It automatically estimates the hyper-parameters α and β . Mallet is used for unlabeled text analysis, particularly topic modeling. Its parameters may be retained as the default value except for topic count. The number of topics defined ahead of time depends on the expected level of topic specialization [65]. Labels are assigned to each of the resulting topics, but

in situations where relatively fewer dimensions are observed, the topics tend to be more general. This is because they represent a broad assortment of terms, which makes it difficult to assign specific labels.

We employed the algorithm on various sets of dimensions and compared the results. We decided to focus on seventy dimensions, as this number gives the opportunity to consider a broader variety of topics without being overly granular. The algorithm produced two output sets per topic. The first output set comprises all the terms of the corpus and the extent to which they are likely to contribute to the topic [65]. The second result set contains all documents in the corpus and the probability that the corresponding topic occurs in the document. For the LDA Mallet model, while we still set the same topic count to twenty, its coherence score improved from 0.32 to 0.41. The LDA Mallet model gives better performance when compared to the regular LDA model. In interpreting the results, the five to twenty most probable terms for each topic are usually examined to identify the degree of commonality and, thus, specify the label of the topic [64]. Since the number of topics can impact the coherence score, experiments were further conducted with varying numbers of topics and the coherence scores compared.

In Figure 7, the various coherence scores based on varying numbers of topics are shown in the line chart. Experiments were carried out with differing number of topics, and the one with the highest coherence score was identified. As shown, the attempt initially has five topics, then ten, fifteen, etc., up to fifty topics. The count (number) of topics refers to the number of nodes (subtopics) used in topic modeling, and in LDA Mallet model.

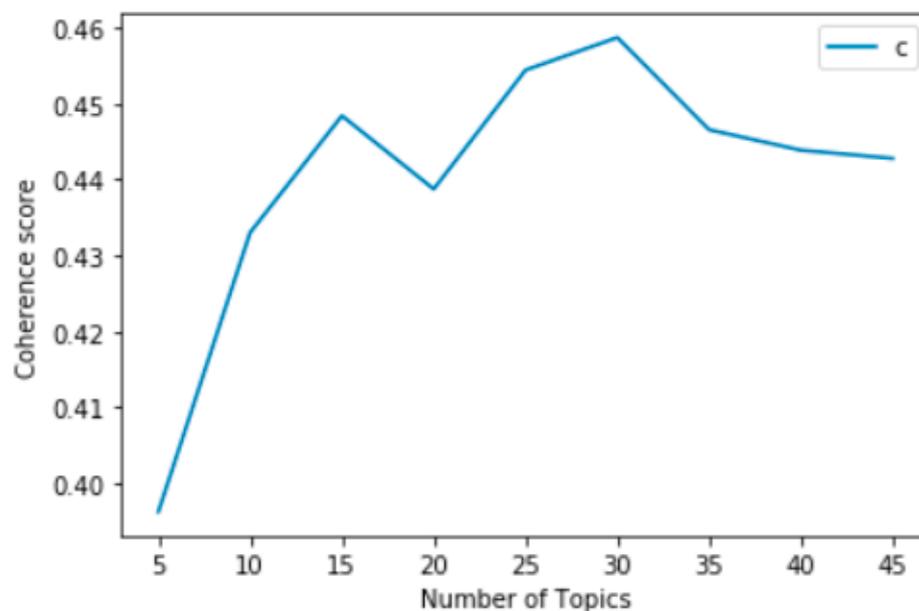


Figure 7. Coherence score for number of topics.

In Figure 8, it is seen that the coherence scores rise gradually with an increase in number of topics, particularly after 5 topics. A maximum is reached at 15 topics. Then, the coherence scores start to decline. At 25 topics, the coherence scores are higher than for 15 topics, however, too many topics will lead to overlapping bubbles.

Figure 9 is a visualization of the distribution of the topics. The numbers in the bubbles in Figures 9 and 10 correspond to the fifteen nodes (sub-topics) identified in our analysis. Therefore, each bubble represents a node that we chose as a sub-topic. The larger the bubble, the more prevalent the topic. The most frequent terms related to data breaches in all the fifteen nodes are “contract”, “card”, “credit”, “consumer”, “employee”, “counsel”, “security”, and “product”.

Number of Topics = 5 has Coherence Value of 0.3963
 Number of Topics = 10 has Coherence Value of 0.433
 Number of Topics = 15 has Coherence Value of 0.4484
 Number of Topics = 20 has Coherence Value of 0.4387
 Number of Topics = 25 has Coherence Value of 0.4544
 Number of Topics = 30 has Coherence Value of 0.4587
 Number of Topics = 35 has Coherence Value of 0.4465
 Number of Topics = 40 has Coherence Value of 0.4438
 Number of Topics = 45 has Coherence Value of 0.4428

Figure 8. Coherence score for number of topics iteration.



Figure 9. General distribution of LDA Mallet 15 topics model.

Figure 10 displays the inter-topic distance map with the topic model on the left containing the fifteen nodes, and the bar chart on the right. Each of the fifteen nodes has certain meaningful keywords. The bars show the frequency of the keyword in the overall corpus. When we select a topic (bubble), the keywords that are most prevalent in that topic will be highlighted in the bars. Taking node 2 as an example, the node refers to the groups of cases that have closely related keywords as topics. The figure shows that when node 2 is selected, the red bars highlight the occurrence of keywords for node 2 compared to the total. Keywords with extended red bars can be considered true keywords that only appear in or are concentrated in the corresponding node. For example, node 2 focuses on “future, identity_theft, speculative, hacker”. Likewise, node 1 on “duty, negligence, deceptive, unfair”; node 4 on “fcra (Fair Credit Reporting Act), receipt, intangible”; node 5 on “contract, insurance, plan, coverage, anthem”; node 6 on “settlement, counsel, fund, award”; node 7 on “cybersecurity, national_union, stock, misleading, share”, node 8 on “product, certification, transfer, vehicle, warranty”; node

9 on “medical, employee, employer, patient, treatment”; node 10 on “phone, call, device, telephone, tcpa (Telephone Consumer Protection Act)”; node 11 on “card, merchant, bank, visa, financial_institution, assessment, authorization”; node 12 on “counsel”, “court”, “jurisdiction”, “litigation”, “party”, “settlement”; node 13 on “election, official, county, voter, board”; node 14 on “confidential, privilege, sanction”; and node 15 on “debt, collection, fdcpa (Fair Debt Collection Practices Act), antitrust, loan”. The nodes have their keyword groups, and in general, among the 698 cases, they could be separated into key data breach-related sub-topic groups including “identity hack security”, “consumer information with fcqa”, “medical/insurance information”, “stock cybersecurity”, “product certification and warranty”, “credit card and bank”, “data breach litigation”, “voting data breach”, and “information leak during debt collection”.

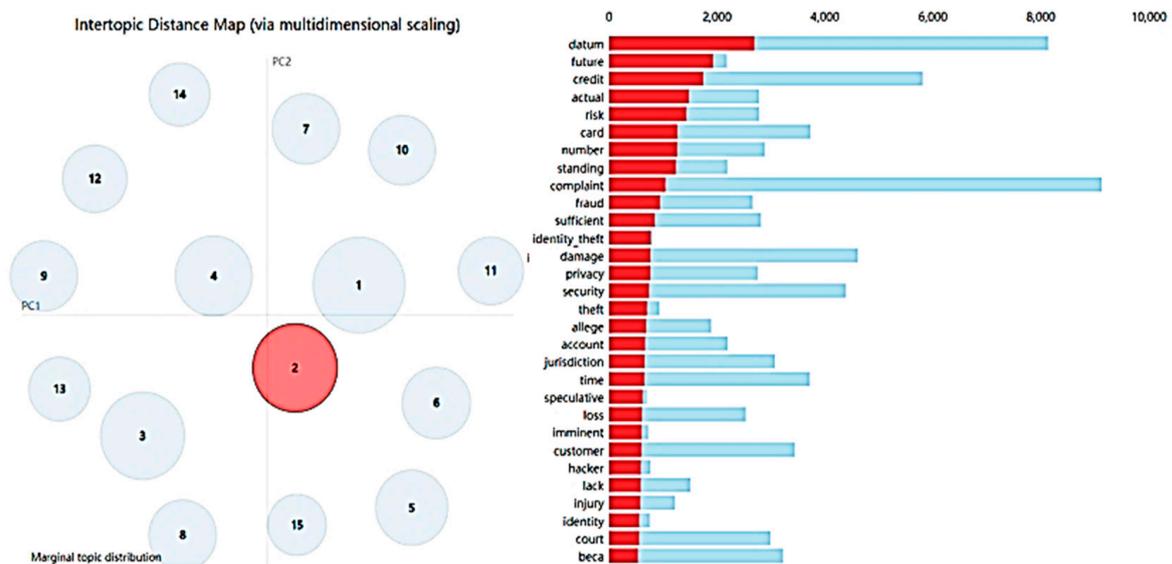


Figure 10. Example node 2 in the inter-topic distance map—distribution of keywords in LDA Mallet node example: top thirty most relevant terms of node 2.

In addition to being visually represented in the topic modeling, the keywords could be displayed directly. The principle of topic modeling is to find the topic count that contributes the highest percentage contribution in a case document. First, we extract the keywords in each sentence and append the frequency to a case. In reversal, when the topic for each sentence is confirmed, we could use Pandas “group by” function to find the representative case for each sub-topic. For example, the most frequent keywords set could apply to 160 of 698 cases, namely, “settlement”, “member”, “counsel”, “fee”, “party”, “rule”, “litigation”, “agreement”, “notice”, and “issue”.

To derive the exact topic distribution for the sample, the topic distribution could be calculated as $\text{topic_counts} / \text{sum of topic_counts}$. Figure 11 shows the top five frequent keyword sets with the topic count set to eight. However, the result exhibits the most frequently occurring keywords instead of the keywords occurring frequently in certain nodes only.

Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
1.0	0.0 settlement, member, counsel, fee, party, rule, litigation, agreement, notice, issue	160.0	0.2292
3.0	1.0 credit, consumer, violation, complaint, actual, future, standing, privacy, concrete, risk	154.0	0.2206
4.0	2.0 evidence, ment, record, discovery, employee, order, request, computer, company, datum	154.0	0.2206
2.0	1.0 credit, consumer, violation, complaint, actual, future, standing, privacy, concrete, risk	120.0	0.1719
0.0	1.0 credit, consumer, violation, complaint, actual, future, standing, privacy, concrete, risk	110.0	0.1576

Figure 11. Top 5 sub-topic distribution.

In addition to the overall topic modeling distribution, we could also analyze what each case is about. To achieve this, we identify the topic number that contributes the highest percentage in each case. For example, we can see the keywords “credit, consumer, violation, complaint, actual, future, standing, privacy, concrete, risk”, occupy 58.53% keywords in the case “064—Hutton v Nat’l Bd of Exam’rs in Optometry Inc”; while the keywords related to settlement: “settlement, member, counsel, fee, party, rule, litigation, agreement, notice, issue”, occupy 40.91% keywords in the case “036—Bliss And Glennon Inc v Ashley” in Figure 12. With this method, in the future, we can zoom in on any case in more detail.

Topic_Num	Topic_Perc_Contrib	Keywords	Keywords
2	2.0	0.8776	evidence, ment, record, discovery employee, order, request, computer, company, datum [075 - Breiterman v United States Capitol Police.pdf3/1/2019, for educational only, breiterman v. united s capitol police, 323 f.r.d. 36 (2017) 99 fed.r.serv.3d 45, 323 f.r.d. united s district court, district of columbia., jodi breiterman, ., v., united s capitol police, ., civil no. 16-0893 (tjkr/mm), , signed 11/07/2017, synopsis background: emplo brought against the united s capitol police (uscsp) alleging gender discrimination and retaliation in violation of title vii and vi
4	4.0	0.8385	contract, datum, security, service, duty, damage, customer, complaint, loss, card [087 - In re VTech Data Breach Litigation.pdf3/1/2019, for educational only, in re vtech data litigation, slip copy (2018) 95 ucc rep.serv.2d 861, united s district court, illinois, eastern division, 2018 wl 1863953, in re vtech data litigation, nos. 15 cv 10889, 15 cv 10891, 15 cv 11620, and 15 cv 11885, , signed 04/18/2018, manish s. sr united s district judge, memorandum opinion and order, *1, vtech electronics north america, llc manuares and markets digital learning toys
0	0.0	0.8261	settlement, member, counsel, fee, party, rule, litigation, agreement, notice, issue [048 - In re Countrywide Financial Corp Customer Data Sec Breach Litigation.pdf3/1/2019, for educational only, in re countrywide financial corp. customer data sec. . . reported in . . . 2010 wl 3341200, only the west citation is currently available., this decision was reviewed by west editorial staff and not assigned editorial enhancemer united s district court, w.d. kentucky, louisville division., in re countrywide financial corp. customer data security litigation., this relates
3	3.0	0.8098	complaint, rule, jurisdiction, letter, section, relief, order, educational, violation, notice [082 - Mathias v York County.pdf3/1/2019, for educational only, mathias v. york county, not reported in fed. supp. (2017), 2017 wl 770610, only the west citation is curre available. united s district court, m.d. pennsylvania., shahnawaz m. mathias, . v., york county, et al., s, no. 1:16-cv-01338, , d.02/28/2017, attorneys and firms, royce l. mo thomas j. weber, camille a. howlett, kathryn e. peters, goldberg, katzman p.c., harrisburg, pa, for ., glenn j. smith, york county sol
1	1.0	0.7970	credit, consumer, violation, complaint, actual, future, standing, privacy, concrete, risk [021 - Brooks v Hualalai Investors LLC.pdf3/1/2019, for educational only, brooks v. hualalai investors, llc, slip copy (2017), 2017 wl 8233902, only the west citation is curre available., united s district court, d. hawai'i., steven brooks, individually and on behalf of all similarly situated individuals, ., hualalai investors, llc, a deare limited liab company, et al., s., v., civil no. 17-00364 jms-kjm, , signed 10/30/2017, attorneys and firms, alex p. katofsky, daniel f. gair

Figure 12. Example of sub-topic in the first five law case files.

4.3. K-Means and Document Similarity

In this section, the core methods of the word2vec model, K-Means, and TF-IDF are utilized to analyze the cases. The TF-IDF (term frequency-inverse document frequency) method was applied in the computation of the weight of each term to denote its importance in a document. This is an information retrieval method that typically assigns weight to a term’s frequency (TF) and its inverse document frequency (IDF). Each term is given the two scores, respectively. The weight of the term is then calculated as a product of these scores. The K-Means clustering model was applied to elicit the key data breach concepts. K-Means clustering is a widely used ML algorithm deployed typically to classify cases based on similarity measures (i.e., the distance between cases). It is typically used in pattern recognition and classification problems.

This was followed by classification with the KNN classifier (a supervised ML algorithm), wherein the data were split into two parts to measure the success of the classifier (training and testing). Prior to the application of ML methods on the text, bigram, and document similarities were generated. This is because keywords by themselves do not contribute very much to a deeper understanding of the nature and dimensions of data breach litigation cases. Therefore, we next examined the co-occurrence of words. In linguistics, co-occurrence characterizes the likelihood of occurrence of two terms in a particular sequence, parallel to each other within a large corpus of data. In this interpretation, it is used as an indicator of the semantic closeness of terms [68]. This model gives insight as to which issues are related. TF-IDF is the key method that is implemented to generate similar documents. Word2vec is applied and tuned accordingly to vectorize the text. This is fed into the K-Means model to cluster. The LDA Mallet is applied to retrieve topic words for each cluster.

Figure 13, a bar chart, shows the top thirty bigrams that come together in the dataset frequently. Patterns in this analysis display frequently discussed topics in our corpus. For example, “data breach”, “breach contract”, “data security”, and “credit card”, are dominant words appearing together in the corpus.

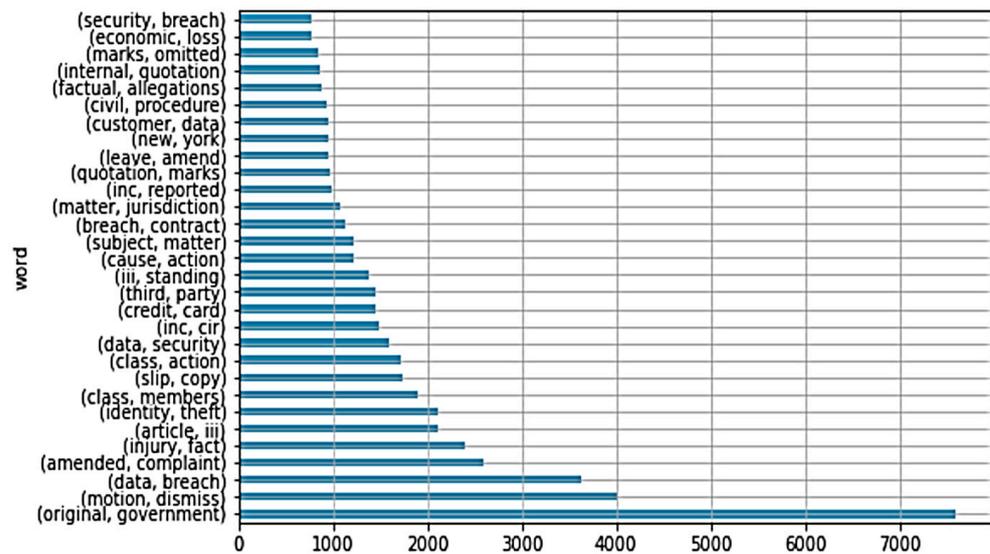


Figure 13. Top 30 words in bigrams.

Figure 14 shows the top thirty words in trigrams, displaying the frequency of three words bundled together in the corpus. Notably, “customer data security”, “social security number”, and “risk identity theft”, are words that come together frequently in the corpus. Additionally, “federal rules civil”, and “economic loss doctrine”, are also frequently appearing, indicating which legal doctrines are frequently discussed and inferred.

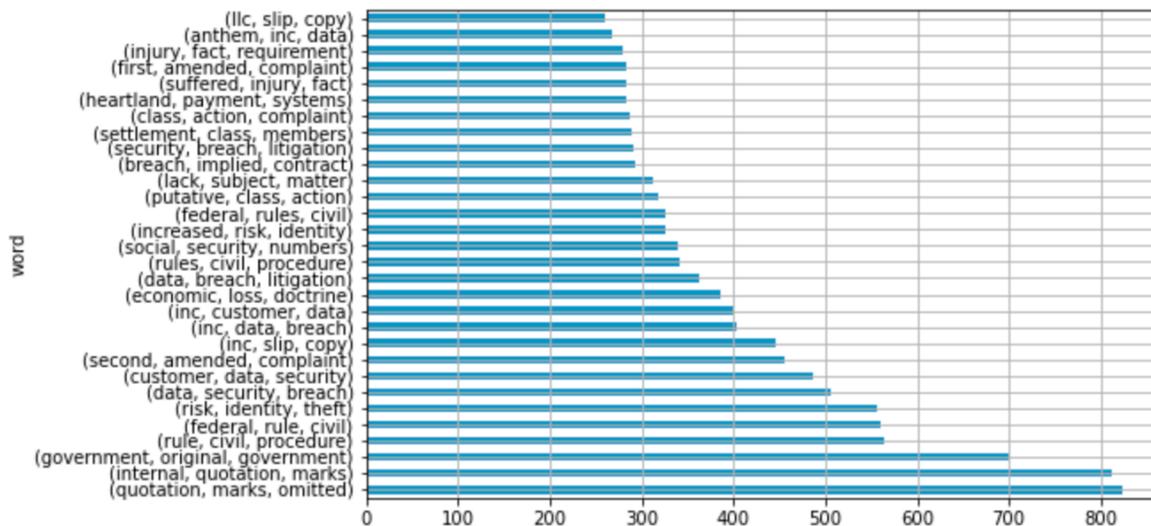


Figure 14. Top 30 words in trigrams.

We also unearthed the key statutes that were most cited in the corpus of cases by implementing Spacy’s rule-based matching and using the keyword “Act”. To better represent the result, we manually checked and retrieved the top 20 Acts (Figure 15). Notably, the top statutes involved “privacy”, “fair credit reporting”, and “health information technology”.

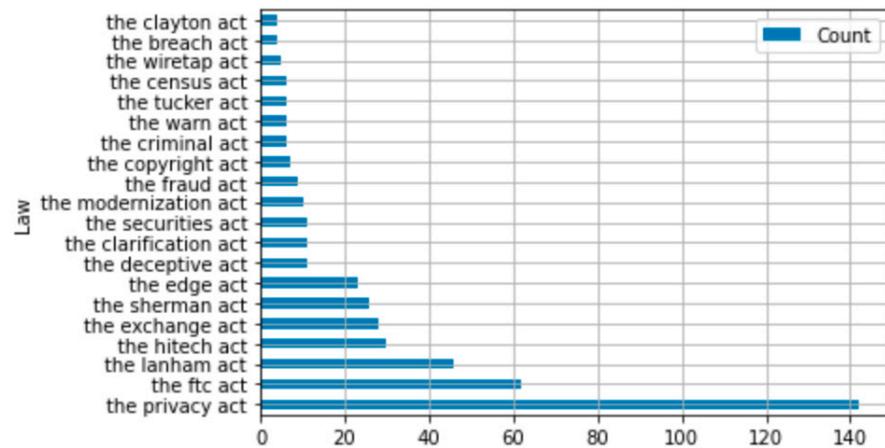


Figure 15. Top 20 statutes.

Figure 16 displays the top twenty previously decided cases mentioned in our documents. In the process of deciding a case, prior cases (precedents) are cited. These cases discussed breaches in terms of “contract”, “confidentiality”, and others. For example, “Clapper v. Amnesty” is about the government electronic surveillance, “Data Corporation v. Realsource” included the issue of data misuse that affected the business profits.

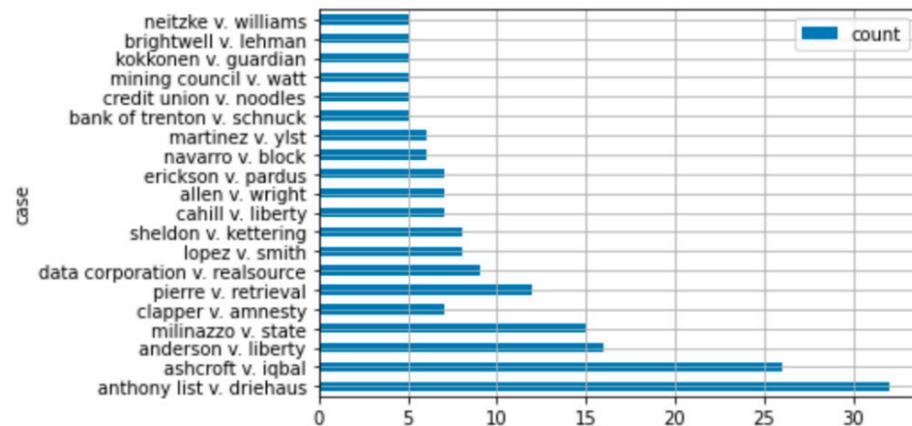


Figure 16. Top 20 litigation cases cited.

Figure 17 displays a document similarity example. We implemented the TF-IDF technique to retrieve the top five most similar documents vis-à-vis a given document. The TF-IDF method can compute a value for each word in the corpus. We calculated the cosine similarity between the selected document and the other documents in the corpus to identify the similarities. The goal is to input a selected document and the function will compute the top five most similar documents based on the similarity score. The result will contain the title of the selected original document, the other five similar documents’ titles, and the similarity score, respectively. Next, we vectorize our textual data and apply clustering algorithms. To input words into the machine learning models, it is necessary to vectorize words based on their linguistic context so that the model can understand the words accordingly. Many tasks use the well-known but simplistic method of the bag of words (BOW) (e.g., TF-IDF), but outcomes will be of low quality since BOW has problems of loss of word order, difficulty of semantic context, etc.

```

Original document :
061 - In re Yahoo! Inc Customer Data Security Breach Litigation.pdf

Most similar document:
-----
064 - Hutton v Nat'l Bd of Exam'rs in Optometry Inc.pdf
TF-IDF Score:
0.39041397520223314
-----
036 - Bliss And Glennon Inc v Ashley.pdf
TF-IDF Score:
0.4679064018273199
-----
019 - Chambliss v Carefirst Inc.pdf
TF-IDF Score:
0.5302504448270196
-----
061 - In re Yahoo! Inc Customer Data Security Breach Litigation.pdf
TF-IDF Score:
0.3226762992622215
-----
041 - Duqum v Scottrade Inc.pdf
TF-IDF Score:
0.4395558888237657
    
```

Figure 17. Document similarity.

Word2vec is implemented to vectorize words. Word2vec is a technique that trains a shallow neural network on individual words in a text and is given words in the vicinity as the label to predict. Prior to the vectorization of the documents, the scraped, raw summary section was transformed into plain tokens through the elimination of punctuation, standard stop words, and lemmatization used in Regex and nltk. Moreover, we loaded a 150-dimensional word vector in Gensim’s Word2vec model to vectorize our filtered tokens. We obtained a dense vector for each word in our corpus, but it would be better to obtain the document-level embeddings from each of the words present in each document.

Therefore, we applied the strategy to average out the word embeddings for each word in a document. In this way, each document has one embedding with 150 dimensions. We were able to generate document features for our corpus preparing for clustering the documents. The goal of clustering is to explore the potential to group documents together. The K-Means algorithm was applied to generate the clusters. After the K-Means model was trained, the most central ten documents in each cluster were generated. Next, to retrieve the prominent keywords for each cluster, we applied the LDA model to retrieve the top fifteen topic words in each cluster. Figure 18 shows the number of documents in each cluster and Figure 19 is a group of words for each cluster.

Cluster	Cluster one	Cluster two	Cluster three	Cluster four
Num of Doc	190	174	170	164

Figure 18. Number of documents in cluster.

point [13,43]. By analyzing the relevant cases, researchers are bringing attention to data breaches, hacking, and cybersecurity issues. Therefore, data breach litigation cases are a key source of information on data breaches. By extrapolation, the various stakeholders, including activists, management, the government, NGOs, and global agencies, can deepen their understanding of data breach phenomena.

In this vein, our exploratory study uses a variety of machine learning-based methods for text analytics to analyze and elicit key data breach information from 698 cases spanning a period of three years. The research capitalizes on advances in information processing technology in extracting insight from large corpora of text, a process that was only entrusted to manual study and subjective evaluation in the past. In the process, and in line with our research question, we identified four clusters and fifteen sub-topics related to data breaches, representing the main concerns. Our analysis shows stakeholders (e.g., litigants) are concerned about major topics related to identity theft, hacker, negligence, FCRA (Fair Credit Reporting Act), cybersecurity, insurance, phone device, TCPA (Telephone Consumer Protection Act), credit card, merchant, privacy, and others. The topics fall into the four major clusters, namely, “phone scam”, “cybersecurity”, “identity theft”, and “business data breach”.

Remarkably, the identification of these topics is consistent with results from other studies related to data breaches, identity theft, and cybersecurity. Analyzing data breach litigation cases allows a window into the landscape of issues that litigants and other stakeholders are concerned about. Analyzing data breach litigation cases also helps consumers and companies foresee the evolution of these issues in the public eye. In addition, it helps companies tell whether they are likely to become targets of public scrutiny related to the issues.

6. Conclusions

This exploratory research seeks to understand key themes and legal implications related to data breaches. Using methodologies such as LDA (Latent Dirichlet Allocation), TF-IDF (term frequency-inverse document frequency), and K-Means clustering, we distilled essential topics related to data breaches. These topics are categorized broadly into four main areas: “phone scam”, “cybersecurity”, “identity theft”, and “business data breach”. Furthermore, utilizing tools like Spacy and Regex, we identified frequently referenced statutes and case citations. The insights derived from this research serve as valuable information for stakeholders in shaping their strategies to address and prevent data breaches in the future.

However, there are many other issues that warrant further investigation. For instance, we need to explore the extent to which text documents, such as data breach litigation cases, can be harnessed to predict trends in data breaches and cybersecurity, as well as influence managerial decisions. Trend analysis could be employed to evaluate the originality of cases and the ramifications of judicial rulings over time. Future research could also employ text analytics and machine learning to delve deeper into litigation documents, aiming to identify how data breaches influence legal outcomes. Moreover, examining the reciprocal relationship between the public’s perception of data breach litigations and the influence of these litigations on the public’s viewpoint can be undertaken from the angle of social media through public sentiment analysis.

While leveraging machine learning and text analytics, one faces considerable challenges in ensuring reproducibility and validation. Furthermore, the process of intuitively labeling clusters and interpreting machine learning outcomes often introduces a subjective element to the analysis, potentially influencing the robustness of the results. This underlines the necessity for developing frameworks that can mitigate subjective bias and enhance the objectivity and reliability of the findings. However, we remain confident that the outcomes of our analysis, including the identification of primary topics and the categorization of clusters, adequately reflect the narratives encapsulated in the cases, a

stance reinforced by the existing literature. This underscores a potent methodology to delve into and comprehend the nuances of data breaches.

Nevertheless, this study is not without its limitations. One concern includes the reliability of the documents and the validity of data preparation techniques. The generalizability of the topics may be constrained, given that they are derived from a small sample of cases. They may not characterize overall global efforts and initiatives regarding data breaches and cybersecurity at the macro level. Further, machine learning models are only capable of extracting a limited amount of insight. Additionally, this study used case documents only. Future research may augment insight from litigation cases with other empirical data.

Despite the limitations, our study contributes to policy and research in four ways. First, practitioners and researchers can utilize the results to prioritize data breach initiatives and examine some of the clusters via different lenses of stakeholders. Second, the study demonstrates the efficacy of machine learning and text analytics in understanding and gaining insight into data breach litigation cases to make informed decisions by conducting descriptive, predictive, and prescriptive analytics. Third, considering the paucity of data regarding data breaches, especially at the business/corporate level, extracting such insight from litigation cases is a novel contribution of this paper. Fourth, macro-level analysis sheds light on the broad important issues in data breach.

Future research can continue to explore and apply advanced techniques like deep learning to delve into case analysis for additional content analysis. For example, prescriptive analytics can be investigated to not only predict outcomes but also suggest impacts and potential strategies. Other avenues for future research include executing cross-industry or inter-state comparisons, exploring global differences, and examining the costs and benefits of litigation versus settlement that impact corporate data breaches and their correlation to company performance. The sophisticated utilization of advanced techniques, such as artificial intelligence and deep learning, promises to accelerate the process of gaining insight from legal cases associated with data breaches. Moreover, applying discovery analytics on the resolutions may shed light on innovation and new product ideas.

As we look ahead, incorporating deep learning in the analysis of legal case topics has the potential to streamline the process of gaining insights from textual data. Additionally, the integration of artificial intelligence is expected to play a pivotal role in predicting judicial outcomes, potentially unveiling tendencies and patterns in judges' rulings. The convergence of law and data science promises to deepen our understanding of the data breach phenomenon, presenting a multidimensional view that integrates legal intricacies with data-driven insights. This combination holds the potential to fundamentally transform our approach to cybersecurity and data protection in the forthcoming era.

Author Contributions: Conceptualization, D.M., W.R., A.S., and V.R.; methodology, D.M., W.R., A.S., and V.R.; software, D.M., W.R., A.S., and V.R.; validation, D.M., W.R., A.S., and V.R.; formal analysis, D.M., W.R., A.S., and V.R.; investigation, D.M., W.R., A.S., and V.R.; resources, D.M., W.R., A.S., and V.R.; data curation, D.M., W.R., A.S., and V.R.; writing, D.M., W.R., A.S., and V.R.; writing—review and editing, D.M., W.R., A.S., and V.R.; visualization, D.M., W.R., A.S., and V.R.; supervision, D.M., W.R., A.S., and V.R.; project administration, D.M., W.R., A.S., and V.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheng, L.; Liu, F.; Yao, D. Enterprise data breach: Causes, challenges, prevention, and future directions. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *7*, e1211. [[CrossRef](#)]
2. Liu, Q.; Li, P.; Zhao, W.; Cai, W.; Yu, S.; Leung, V.C. A survey on security threats and defensive techniques of machine learning: A data driven view. *IEEE Access* **2018**, *6*, 12103–12117. [[CrossRef](#)]

3. Schlackl, F.; Link, N.; Hoehle, H. Antecedents and consequences of data breaches: A systematic review. *Inf. Manag.* **2022**, *59*, 103638. [CrossRef]
4. IBM. Cost of a Data Breach Report. 2023. Available online: <https://www.ibm.com/downloads/cas/E3G5JMBP> (accessed on 28 September 2023).
5. PwC. PwC's 23rd Annual Global CEO Survey. 2023. Available online: <https://www.pwc.com/gx/en/issues/c-suite-insights/ceo-survey-2023.html> (accessed on 28 September 2023).
6. Dhillon, G.; Smith, K.; Dissanayaka, I. Information systems security research agenda: Exploring the gap between research and practice. *J. Strateg. Inf. Syst.* **2021**, *30*, 101693. [CrossRef]
7. Layton, R.; Watters, P.A. A methodology for estimating the tangible cost of data breaches. *J. Inf. Secur. Appl.* **2014**, *19*, 321–330. [CrossRef]
8. Sherstobitoff, R. Anatomy of a data breach. *Inf. Secur. J. A Glob. Perspect.* **2008**, *17*, 247–252. [CrossRef]
9. Watters, P.A. *Cyber Security: Concepts and Cases*; CreateSpace Independent Publishing Platform: North Charleston, SC, USA, 2012.
10. Irwin, L. The 6 Most Common Ways Data Breaches Occur. 2020. Available online: <https://www.itgovernance.eu/blog/en/the-6-most-common-ways-data-breaches-occur> (accessed on 28 September 2023).
11. Wang, P.; Johnson, C. Cybersecurity incident handling: A case study of the Equifax data breach. *Issues Inf. Syst.* **2018**, *19*, 3.
12. Wang, P.; Park, S.-A. Communication in Cybersecurity: A Public Communication Model for Business Data Breach Incident Handling. *Issues Inf. Syst.* **2017**, *18*, 2.
13. Romanosky, S.; Hoffman, D.; Acquisti, A. Empirical analysis of data breach litigation. *J. Empir. Leg. Stud.* **2014**, *11*, 74–104. [CrossRef]
14. Sanzgiri, A.; Dasgupta, D. Classification of insider threat detection techniques. In Proceedings of the 11th Annual Cyber and Information Security Research Conference, Oak Ridge, TN, USA, 5–7 April 2016; pp. 1–4.
15. Congressional Research Service. 2015. Available online: <https://crsreports.congress.gov/> (accessed on 1 November 2023).
16. CNN. Yahoo Says 500 Million Accounts Stolen. 2016. Available online: <https://money.cnn.com/2016/09/22/technology/yahoo-data-breach/> (accessed on 1 November 2023).
17. McAfee. Grand Theft Data. 2017. Available online: <https://www.mcafee.com> (accessed on 1 November 2023).
18. Greenberg, A. More than Half of Corporate Breaches Go Unreported, according to Study. 2013. Available online: <https://www.scmagazine.com/news/more-than-half-of-corporate-breaches-go-unreported-according-to-study> (accessed on 1 November 2023).
19. Huq, N. Follow the data: Dissecting data breaches and debunking myths. *TrendMicro Res. Pap.* **2015**.
20. McGee Kolbasuk, M. Why Data Breaches go Unreported. 2014. Available online: <https://www.bankinfosecurity.com/health-data-breaches-go-unreported-a-6804> (accessed on 1 November 2023).
21. Privacy Rights Clearinghouse. Data Breaches Chronology. 2023. Available online: <https://privacyrights.org/data-breaches> (accessed on 28 September 2023).
22. Anderson, R.; Barton, C.; Böhme, R.; Clayton, R.; Van Eeten, M.J.; Levi, M.; Moore, T.; Savage, S. Measuring the cost of cybercrime. In Proceedings of the 11th Workshop on the Economics of Information Security (WEIS), Washington, DC, USA, 11–12 June 2013; pp. 265–300.
23. U.S. News. Equifax Breach Could Have ‘Decades of Impact’. 2017. Available online: <https://www.usnews.com/news/articles/2017-09-08/equifax-breach-could-have-decades-of-impact-on-consumers> (accessed on 1 November 2023).
24. Saleem, H.; Naveed, M. SoK: Anatomy of data breaches. *Proc. Priv. Enhancing Technol.* **2020**, *2020*, 153–174. [CrossRef]
25. Bielinski, C. 2018 Trustwave Global Security Report. 2018. Available online: <https://www.trustwave.com/en-us/resources/library/documents/2018-trustwave-global-security-report/> (accessed on 1 November 2023).
26. Manworren, N.; Letwat, J.; Daily, O. Why you should care about the Target data breach. *Bus. Horiz.* **2016**, *59*, 257–266. [CrossRef]
27. Rashid, A.; Ramdhany, R.; Edwards, M.; Kibirige Mukisa, S.; Ali Babar, M.; Hutchison, D.; Chitchyan, R. *Detecting and Preventing Data Exfiltration*; Lancaster University: Lancaster, UK, 2014.
28. Collins, J.D.; Sainato, V.A.; Khey, D.N. Organizational data breaches 2005-2010: Applying SCP to the healthcare and education sectors. *Int. J. Cyber Criminol.* **2011**, *5*, 794.
29. Posey Garrison, C.; Ncube, M. A longitudinal analysis of data breaches. *Inf. Manag. Comput. Secur.* **2011**, *19*, 216–230. [CrossRef]
30. Ayyagari, R. An exploratory analysis of data breaches from 2005-2011: Trends and insights. *J. Inf. Priv. Secur.* **2012**, *8*, 33–56. [CrossRef]
31. Khey, D.N.; Sainato, V.A. Examining the correlates and spatial distribution of organizational data breaches in the United States. *Secur. J.* **2013**, *26*, 367–382. [CrossRef]
32. Zadeh, A. Characterizing Data Breach Severity: A Data Analytics Approach. AMCIS. 2022. Available online: https://aisel.aisnet.org/treos_amcis2022/19 (accessed on 1 November 2023).
33. Hammouchi, H.; Cherqi, O.; Mezzour, G.; Ghogho, M.; El Koutbi, M. Digging deeper into data breaches: An exploratory data analysis of hacking breaches over time. *Procedia Comput. Sci.* **2019**, *151*, 1004–1009. [CrossRef]
34. Shu, X.; Tian, K.; Ciambone, A.; Yao, D. Breaking the target: An analysis of target data breach and lessons learned. *arXiv* **2017**, arXiv:1701.04940.
35. Smith, T.T. Examining Data Privacy Breaches in Healthcare. Ph.D. Thesis, Walden University, Minneapolis, MN, USA, 2016.
36. Holtfreter, R.E.; Harrington, A. Data breach trends in the United States. *J. Financ. Crime* **2015**, *22*, 242–260. [CrossRef]

37. Neto, N.N.; Madnick, S.; Paula, A.M.G.D.; Borges, N.M. Developing a global data breach database and the challenges encountered. *J. Data Inf. Qual. (JDIQ)* **2021**, *13*, 1–33. [[CrossRef](#)]
38. McLeod, A.; Dolezel, D. Cyber-analytics: Modeling factors associated with healthcare data breaches. *Decis. Support Syst.* **2018**, *108*, 57–68. [[CrossRef](#)]
39. Algarni, A.M.; Malaiya, Y.K. A consolidated approach for estimation of data security breach costs. In Proceedings of the 2016 2nd International Conference on Information Management (ICIM), London, UK, 7–8 May 2016; pp. 26–39.
40. Kafali, Ö.; Jones, J.; Petruso, M.; Williams, L.; Singh, M.P. How good is a security policy against real breaches? A HIPAA case study. In Proceedings of the 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE), Buenos Aires, Argentina, 20–28 May 2017; pp. 530–540.
41. Sen, R.; Borle, S. Estimating the contextual risk of data breach: An empirical approach. *J. Manag. Inf. Syst.* **2015**, *32*, 314–341. [[CrossRef](#)]
42. Hall, A.A.; Wright, C.S. Data security: A review of major security breaches between 2014 and 2018. *Fed. Bus. Discip. J.* **2018**, *6*, 50–63.
43. Romanosky, S. Examining the costs and causes of cyber incidents. *J. Cybersecur.* **2016**, *2*, 121–135. [[CrossRef](#)]
44. Goode, S.; Hoehle, H.; Venkatesh, V.; Brown, S.A. User compensation as a data breach recovery action. *MIS Q.* **2017**, *41*, 703–728. [[CrossRef](#)]
45. As Data Breach Class Actions Arise. New York Law Journal. 2023. Available online: <https://www.law.com/newyorklawjournal/?slreturn=20231005012616> (accessed on 1 November 2023).
46. 2021 Year in Review: Data Breach and Cybersecurity Litigations. The National Law Review. 2021. Available online: <https://www.privacyworld.blog/2021/12/2021-year-in-review-data-breach-and-cybersecurity-litigations/> (accessed on 1 November 2023).
47. Black, M. HCA Data Breach Class Action Lawsuit May Include 11 Million; Mission Patients Notified. Asheville Citizen Times. 2023. Available online: <https://www.citizen-times.com/story/news/local/2023/08/29/hca-data-breach-class-action-lawsuit-may-represent-11-million-patients/70699685007/> (accessed on 1 November 2023).
48. Yenouskas, J.; Swank, L. Emerging Legal Issues in Data Breach Class Actions. 2018. Available online: https://www.americanbar.org/groups/business_law/resources/business-law-today/2018-july/emerging-legal-issues-in-data-breach-class-actions/ (accessed on 1 November 2023).
49. Hill, M.; Swinhoe, D. The 15 Biggest Data Breaches of the 21st Century. 2021. Available online: <https://www.csoonline.com/article/534628/the-biggest-data-breaches-of-the-21st-century.html> (accessed on 1 November 2023).
50. Bellamy, F.D. Data Breach Class Action Litigation and Changing Legal Landscape. 2022. Available online: <https://www.reuters.com/legal/legalindustry/data-breach-class-action-litigation-changing-legal-landscape-2022-06-27/> (accessed on 1 November 2023).
51. Khan, A.; Baharudin, B.; Lee, L.H.; Khan, K. A review of machine learning algorithms for text-documents classification. *J. Adv. Inf. Technol.* **2010**, *1*, 4–20.
52. Landmann, J.; Zuell, C. Identifying events using computer-assisted text analysis. *Soc. Sci. Comput. Rev.* **2008**, *26*, 483–497. [[CrossRef](#)]
53. Ford, J.M. Content analysis: An introduction to its methodology. *Pers. Psychol.* **2004**, *57*, 1110.
54. Raghupathi, V.; Ren, J.; Raghupathi, W. Studying public perception about vaccination: A sentiment analysis of tweets. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3464. [[CrossRef](#)]
55. Raghupathi, V.; Zhou, Y.; Raghupathi, W. Legal decision support: Exploring big data analytics approach to modeling pharma patent validity cases. *IEEE Access* **2018**, *6*, 41518–41528. [[CrossRef](#)]
56. Raghupathi, V.; Zhou, Y.; Raghupathi, W. Exploring big data analytic approaches to cancer blog text analysis. *Int. J. Healthc. Inf. Syst. Inform. (IJHISI)* **2019**, *14*, 1–20. [[CrossRef](#)]
57. Ren, J.; Raghupathi, V.; Raghupathi, W. Understanding the dimensions of medical crowdfunding: A visual analytics approach. *J. Med. Internet Res.* **2020**, *22*, e18813. [[CrossRef](#)]
58. Székely, N.; Vom Brocke, J. What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. *PLoS ONE* **2017**, *12*, e0174807. [[CrossRef](#)]
59. Zhou, Y.; Wang, X.; Yuen, K.F. Sustainability disclosure for container shipping: A text-mining approach. *Transp. Policy* **2021**, *110*, 465–477. [[CrossRef](#)]
60. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [[CrossRef](#)]
61. Graham, S.; Weingart, S.; Milligan, I. Getting Started with Topic Modeling and MALLET. The Editorial Board of the Programming Historian. 2012. Available online: <https://uwspace.uwaterloo.ca/handle/10012/11751> (accessed on 1 November 2023).
62. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [[CrossRef](#)]
63. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **2016**, *5*, 1–22. [[CrossRef](#)]
64. Crain, S.P.; Zhou, K.; Yang, S.-H.; Zha, H. Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. *Min. Text Data* **2012**, 129–161.

65. Krestel, R.; Fankhauser, P. Tag recommendation using probabilistic topic models. *ECML PKDD Discov. Chall.* **2009**, *2009*, 131.
66. Debortoli, S.; Müller, O.; Junglas, I.; Vom Brocke, J. Text mining for information systems researchers: An annotated topic modeling tutorial. *Commun. Assoc. Inf. Syst. (CAIS)* **2016**, *39*, 7. [[CrossRef](#)]
67. Syed, S.; Spruit, M. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In Proceedings of the 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan, 18 January 2018; pp. 165–174.
68. Yi, Y.; Liu, L.; Li, C.H.; Song, W.; Liu, S. Machine learning algorithms with co-occurrence based term association for text mining. In Proceedings of the 2012 Fourth International Conference on Computational Intelligence and Communication Networks, Mathura, India, 3–5 November 2012; pp. 958–962.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.